MACHINE LEARNING FOR GEOLOGICAL MAPPING: ALGORITHMS AND APPLICATIONS

UNIVERSITY OF TASMANIA

MATTHEW J. CRACKNELL

BSc (Hons)

ARC Centre of Excellence in Ore Deposits (CODES)

School of Physical Sciences (Earth Sciences)

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Tasmania

May, 2014

Did you ever fly a kite in bed?

Did you ever walk with ten cats on your head?

Did you ever milk this kind of cow?

Well, we can do it.

We know how.

If you never did you should.

These things are fun and fun is good.

Dr. Seuss

DECLARATION OF ORIGINALITY

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

AUTHORITY OF ACCESS

This non-published content of the thesis (see below) may be made available for loan and limited copying and communication in accordance with the Copyright Act 1968.

STATEMENT REGARDING PUBLISHED WORK CONTAINED IN THESIS

Matthew J. Cracknell

May 2014

STATEMENT OF CO-AUTHORSHIP

The following people and institutions contributed to the publication of work undertaken as part of this thesis:

Matthew James Cracknell, ARC Centre of Excellence in Ore Deposits (CODES), School of Earth Sciences, University of Tasmania = **Candidate**

Anya Marie Reading, ARC Centre of Excellence in Ore Deposits (CODES), School of Earth Sciences, University of Tasmania = **Author 1**

Andrew William McNeill, Mineral Resources Tasmania, Department of Infrastructure Energy & Resources (DIER) = **Author 2**

Author details and their roles:

Paper 1, 'Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information':

Located in Chapter 4

Candidate was the primary author and with Author 1 contributing to its development, refinement and presentation.

Paper 2, 'The upside of uncertainty: Identification of lithology contact zones from airborne geophysics and satellite data using Random Forests and Support Vector Machines':

Located in Chapter 5

Candidate was the primary author and with Author 1 contributing to development, refinement and presentation.

Paper 3, 'Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer–Mt Charter region, Tasmania, using Random Forests™ and Self-Organising Maps':

Located in Chapter 6

Candidate was the primary author and with Author 1 contributing to its refinement and presentation and Author 2 contributing to its formalisation and development.

We the undersigned agree with the above stated "proportion of work undertaken" for each of the above published (or submitted) peer-reviewed manuscripts contributing to this thesis:

Signed: _____ _____

Anya M. Reading	*Jocelyn McPhie*
Supervisor	*Head of School*
School Of Earth Sciences	*School Of Earth Sciences*
University of Tasmania	University of Tasmania

Date: _____ _____

ABSTRACT

Machine learning algorithms are designed to identify efficiently and to predict accurately patterns within multivariate data. They provide analysts computational tools to aid predictive modelling and the interpretation of interactions between data and the phenomena under investigation. The analysis of large volumes of disparate multivariate geospatial data using machine learning algorithms therefore offers great promise to industry and research in the geosciences. Geoscience data are frequently characterised by a restriction in the number and distribution of direct observations, irreducible noise in these data and a high degree of intraclass variability and interclass similarity. The choice of machine learning algorithm, or algorithms and the details of how algorithms are applied must therefore be appropriate to the context of geoscience data. With this knowledge, I aim to employ machine learning as a means of understanding the spatial distribution of complex geological phenomena.

I conduct a rigorous and comprehensive comparison of machine learning algorithms, representing the five general machine learning strategies, for supervised lithology classification applications. I also develop and test a novel method for obtaining robust estimates of the uncertainty associated with machine learning algorithm categorical predictions. The insights gained from these experiments leads to the further development and comparison of new methods for the incorporation of spatial-contextual information into machine learning supervised classifiers.

In using machine learning algorithms for geoscience applications, I have developed best-practice methodologies that address the challenges facing geoscientists for geospatial supervised classification. Guidelines are established that detail the preparation and integration of disparate spatial data, the optimisation of trained classifiers for a given application and the robust statistical and spatial evaluation of outputs. I demonstrate, through a case study in a region that is prospective for economic mineralisation, the combination of supervised and unsupervised machine learning algorithms for the critical appraisal of pre-existing geological maps and formulation of meaningful interpretations of geological phenomena.

The experiments conducted as part of my research confirm the efficacy of machine learning algorithms to generate accurate geological maps representing a variety of terranes. I identify and explore key aspects of the spatial and statistical distributions of geoscience data that affect machine learning algorithm performance. My research clearly identifies Random Forests™ as a good first-choice algorithm for the prediction of classes representing lithologies using commonly available multivariate geological and geophysical data. Furthermore, Random Forests prediction uncertainty is shown to be closely related to ambiguous and/or erroneous classifications and, thus provides a practical means of indicating variable levels of confidence. Spatial-contextual information is best incorporated into machine learning supervised classifiers via the pre-processing of input variables and/or the post-regularisation of classifications. My findings indicate that a trade-off between optimal predictive models and interpretable explanatory models exists, whereby, intuitively interpretable models are not necessarily the most accurate.

The practical application of machine learning algorithms requires the implementation of three key stages: (1) data pre-processing; (2) algorithm training; and (3) prediction evaluation. This methodology provides the foundation for generating accurate and geologically meaningful predictions with minimal user intervention and assists in the formulation of robust interpretations of complex geological phenomena. For example, classifications obtained by Random Forests are useful for critically appraising interpreted geological maps. Clusters produced by Self-Organising Maps indicate the presence of discrete, spatially contiguous and geologically significant sub-classes within individual lithological units, which represent regions of contrasting primary composition and alteration styles. My results may be widely applied to a broad range of practical geoscience challenges such as ore deposit targeting, geo-hazard risk assessment, engineering and construction projects, hydrological and environmental modelling and ecological studies. The applications of machine learning algorithms detailed in this thesis align well with state-of-the-art Big Data online infrastructure and virtual laboratories currently emerging in Australia.

CONTENTS

DECLARATION OF ORIGINALITY III

AUTHORITY OF ACCESS III

STATEMENT REGARDING PUBLISHED WORK CONTAINED IN THESIS III

STATEMENT OF CO-AUTHORSHIP V

ABSTRACT VII

CONTENTS IX

LIST OF TABLES XV

LIST OF FIGURES XVII

LIST OF ABBREVIATIONS XXI

ACKNOWLEDGEMENTS XXIII

CHAPTER 1 – INTRODUCTION 1

1.1. Machine learning **2**

1.2. Geological maps **4**

1.3. Research scope and hypothesis **5**

1.3.1. Major research questions to be addressed 6

1.4. Thesis structure **7**

CHAPTER 2 – MACHINE LEARNING THEORY AND IMPLEMENTATION 9

2.1. Machine learning **9**

2.1.1. Supervised versus unsupervised learning 10

2.2. Supervised classification **10**

2.2.1. Classification strategies 11

2.2.1.1. Statistical learning algorithms 11

2.2.1.2. Instance-based learners 14

2.2.1.3. Logic-based learners 17

2.2.1.4. Support Vector Machines 20

2.2.1.5. Perceptrons 23

2.2.2. Supervised classifier implementation 25

2.2.2.1. Data pre-processing 26

2.2.2.2. Classifier training 27

2.2.2.3. Prediction evaluation 29
2.3. Unsupervised clustering 33
2.3.1. Clustering strategies 33
2.3.1.1. Partitioning algorithms 33
2.3.1.2. Hierarchical algorithms 35
2.3.1.3. Self-Organising Maps 36
2.3.2. Unsupervised clustering implementation 38
2.4. Conclusions 38

CHAPTER 3 – A REVIEW OF MACHINE LEARNING FOR GEOSCIENCE CLASSIFICATION APPLICATIONS 41

3.1. Machine learning non-geoscience applications 41
3.2. Machine learning geoscience applications 44
3.2.1. Classification of 0D data 45
3.2.1. Classification of 1D data 46
3.2.1.1. One temporal dimension 46
3.2.1.2. One spatial dimension 47
3.2.1. Classification of 2D data 51
3.2.1.3. Land cover/vegetation mapping 52
3.2.1.4. Geological mapping 55
Supervised classification 55
Unsupervised clustering 58
Combined supervised and unsupervised methods 60
3.3. Practical machine learning implementation 61
3.3.1. Data 63
3.3.2. Data pre-processing 64
3.3.3. Prediction evaluation 64
3.3.4. Integrated workflow 65
3.4. Conclusions 66

CHAPTER 4 – GEOLOGICAL MAPPING USING REMOTE SENSING DATA: A COMPARISON OF FIVE MACHINE LEARNING ALGORITHMS, THEIR RESPONSE TO VARIATIONS IN THE SPATIAL DISTRIBUTION OF TRAINING DATA AND THE USE OF EXPLICIT SPATIAL INFORMATION 69

4.0. Abstract 69
4.1. Introduction 70
4.1.1. Machine learning for supervised classification 72
4.1.2. Machine learning algorithm theory 73
4.1.2.1. Naïve Bayes 73
4.1.2.2. *k*-Nearest Neighbours 73

4.1.2.3. Random Forests 73
4.1.2.4. Support Vector Machines 74
4.1.2.5. Artificial Neural Networks 74
4.1.3. Geology and tectonic setting 75
4.2. Data 77
4.3. Methods 78
4.3.1. Pre-processing 78
4.3.2. Classification model training 79
4.3.3. Prediction evaluation 79
4.4. Results 79
4.5. Discussion 84
4.5.1. Machine learning algorithms compared 84
4.5.2. Influence of training data spatial distribution 87
4.5.3. Using spatially constrained data 88
4.6. Conclusions 89
4.7. Acknowledgements 90
4.8. Description of supplementary information 91

CHAPTER 5 – THE UPSIDE OF UNCERTAINTY: IDENTIFICATION OF LITHOLOGY CONTACT ZONES FROM AIRBORNE GEOPHYSICS AND SATELLITE DATA USING RANDOM FORESTS AND SUPPORT VECTOR MACHINES 93

5.0. Abstract 93
5.1. Introduction 94
5.1.1. The lithology prediction problem 97
5.1.2. Random Forests 98
5.1.3. Support Vector Machines 99
5.2. Data 101
5.2.1. Tectonic setting and history 101
5.2.2. Data sources 103
5.2.3. Data pre-processing 103
5.3. Methods 103
5.3.1. Training and evaluating algorithms 105
5.3.2. Variance 106
5.4. Results 106
5.5. Discussion 114
5.6. Conclusions 118
5.7. Acknowledgements 119

CHAPTER 6 – MAPPING GEOLOGY AND VOLCANIC-HOSTED MASSIVE SULFIDE ALTERATION IN THE HELLYER–MT CHARTER REGION, TASMANIA, USING RANDOM FORESTS™ AND SELF-ORGANISING MAPS 121

6.0. Abstract **121**
6.1. Introduction **122**
6.1.1. Geological setting 123
6.1.2. Random Forests 128
6.1.3. Self-Organising Maps 130
6.2. Data and Methods **130**
6.2.1. Source data 130
6.2.2. Data sampling 131
6.2.3. Training Random Forests and variable selection 133
6.2.4. Implementing Self-Organising Maps 136
6.3. Results **137**
6.3.1. Geological classification using Random Forests 137
6.3.2. Discrimination of geological sub-classes using Self-Organising Maps 141
6.4. Discussion **144**
6.5. Conclusions **146**
6.6. Acknowledgements **147**

CHAPTER 7 – SPATIAL-CONTEXTUAL MACHINE LEARNING SUPERVISED CLASSIFIERS: LITHOSTRATIGRAPHY CLASSIFICATION EXAMPLE 149

7.0. Abstract **149**
7.1. Introduction **150**
7.1.1. Pre-processing methods 152
7.1.1.1. Focal operators 152
7.1.1.2. Image segmentation 153
7.1.2. Training data selection 154
7.1.3. Post-processing methods 155
7.1.4. Combination methods 155
7.1.5. Study aims 155
7.2. Data **156**
7.2.1. Lithostratigraphy – classification target 156
7.2.2. Geophysical data – input variables 159
7.2.2.1. Pre-processing 160
7.3. Methods **160**
7.3.1. Data sampling 160
7.3.2. Global pixel-based classifiers 162

7.3.3. Spatial-contextual classifiers 162
7.3.3.1. Pre-processing 162
7.3.3.2. Algorithm training 164
7.3.3.3. Post-processing 165
7.3.4. Prediction evaluation 165
7.4. Results 165
7.5. Discussion 173
7.5.1. Spatial-contextual classifiers compared 173
7.5.2. Issues of spatial scale 175
7.5.3. Geological interpretations 176
7.6. Conclusions 177

CHAPTER 8 – SYNTHESIS AND DISCUSSION 179

8.1. Algorithms 179
8.1.1. Supervised classification 179
8.1.1.1. Implementation 180
8.1.1.2. Decision structures 181
8.1.1.3. Accuracy comparison 181
8.1.1.4. Spatial-contextual classifiers 183
8.1.1.5. Prediction uncertainty 184
8.1.2. Unsupervised clustering 185
8.2. Applications 186
8.2.1. Data pre-processing 186
8.2.1.1. Data preparation 187
8.2.1.2. Variable extraction 188
8.2.1.3. Variable selection 189
8.2.2. Classifier training 189
8.2.2.1. Training and test data 190
8.2.2.2. Classifier induction 190
8.2.2.3. Classification post-processing 191
8.2.3. Evaluation and interpretation 192
8.2.3.1. Statistical evaluation 193
8.2.3.2. Interrogating decision structures 194
8.2.3.3. Complementary interpretation 197
8.3. Extended research implications 199
8.3.1. Integrated workflow using R 199
8.3.2. Wider geoscience applications 200
8.3.3. Big Data 202

CHAPTER 9 – CONCLUSIONS 205

REFERENCES 209

APPENDIX A – MACHINE LEARNING ALGORITHM SENSITIVITY TO IMBALANCED CLASS DISTRIBUTIONS 253

A.1. Introduction **253**
A.2. Methods **254**
A.3. Results **256**
A.4. Discussion and Conclusions **259**

APPENDIX B – VARIANCE AND ENTROPY FOR MULTICLASS CLASSIFICATION UNCERTAINTY 261

APPENDIX C – SUPPLEMENTARY INFORMATION 263

C.1. Data **263**
C.2. MLA software and parameters **266**

APPENDIX D – R PACKAGES 269

APPENDIX E – DATA SOURCES AND PRE-PROCESSING 271

APPENDIX F – R CODE AND SCRIPTS 275

README.txt **275**

LIST OF TABLES

Table 2.1 Common distance metrics used to measure the separation distance between samples in multi-dimensional variable space, after Hechenbichler & Schliep (2004) and Kotsiantis (2007). 15
Table 2.2 Common kernel functions for SVM, after Karatzoglou et al. (2006).....22
Table 4.1 Summary of 13 lithological classes within the Broken Hill study region, complied using information from Willis et al. (1983) and Buckley et al. (2002). 76
Table 4.2 MLA specific parameters evaluated during classifier training. Note RF parameters presented indicate those used for all input variables (All Data).....79
Table 4.3 Comparison of MLA cross-validation and T_b accuracy and kappa using all input variables with respect to different numbers of T_a clusters.....83
Table 5.1 Description of input variables: units, resolution; and pre-processing methods.....104
Table 5.2 Uncertainty threshold values for RF and SVM across T_a sample proportions.....109
Table 5.3 Comparison of overall T_b accuracies before and after the elimination of unclassified samples using uncertainty thresholds presented in Table 5.2. 109
Table 5.4 RF and SVM confusion matrices for 0.05 Ta sample proportion classification models.....111
Table 5.5.Comparison of RF and SVM class dependent measures of recall and precision rates and their respective differences as obtained from the 0.05 T_a proportion classification models. 111
Table 6.1 Hellyer–Mt Charter lithological units and their stratigraphic relationships (Corbett & Komyshan 1989; Waters & Wallace 1992).....126
Table 6.2 Pre-processed input datasets (variables) used in this study. 131
Table 6.3 Comparison of parameter selection results of the different stages of variable selection. 135
Table 6.4 T_b confusion matrix for RF predictions.....138
Table 6.5 Comparison of RF T_b recall and precision rates 140
Table 7.1 Summaries of lithostratigraphic classes 157
Table 7.2 Comparison T_b performance statistics for spatial-contextual classifiers 166
Table 7.3 Comparison of the difference between mean T_b accuracy obtained using PR majority focal operators of 3 × 3, 7 × 7 and 11 × 11 neighbourhood (pixel) dimensions and mean T_b accuracy resulting from predictions not utilising majority focal operators (see Table 7.2).....170
Table 8.1 Summaries of the R code and scripts provided in (digital) Appendix F.....200
Table A.1 MLA model parameter values used for “no-information” accuracy prediction (default if possible). 255
Table A.3 Equal and unequal class distributions for c = 2, 3 and 6 classes, the unequal distributions for the 6 class prediction task is taken from the distributions found in real-world data. 256
Table A.2 Training and validation class distribution combinations for trials 1–4.....256
Table A.4 MLA T_b accuracy mean ± one standard deviation (n = 1000) across four trials (Table A.2) of equal and unequal class distributions (Table A.3) for two, three and six classes. 257

LIST OF FIGURES

Figure 1.1 Schematic representation of data inference. 2
Figure 1.2 Schematic representation of the process of deductive and inductive reasoning. 3
Figure 2.1 An example of NB estimated joint (normal) class conditional probability densities for one discretised continuous variable and three classes. 12
Figure 2.2 Schematic diagram of kNN classifications for a binary classification task in 2D variable space. Filled symbols represent predicted class for an unlabelled sample. 14
Figure 2.3 Schematic representation of binary DT classifier architecture. Split nodes partitions inputs based on some threshold (for continuous data) 17
Figure 2.4 SVM schematic diagrams of a) separating hyperplane in 2D variable space, where indicates class a and class b. Filled symbols represent support vectors. b) Kernel transformation example from 1D variable space to 2D kernel space 21
Figure 2.5 Schematic diagram of a) single McCulloch–Pitts neuron. 24
Figure 2.6 Generalised workflow for machine learning supervised classification. 25
Figure 2.7 Confusion (error) matrix for binary classifications showing the relationships between True and False, Positives and Negatives. 30
Figure 2.8 Schematic diagram representing the effect of noise and the bias^2-variance decomposition and its relationship to classifier complexity (under-fitted and over-fitted models) and error 31
Figure 2.9 Example of k-means (k = 3) clustering in 2D variable space. Filled symbols represent cluster centroids and lines indicate boundaries between clusters. 34
Figure 2.10 Example of a dendrogram with five clusters as output from hierarchical clustering, similar clusters reside on the same branch of the dendrogram. 35
Figure 2.11 Schematic diagram of SOM unsupervised clustering algorithm structure, after Bierlein et al. (2008). 37
Figure 4.1 Reference geological map, after Buckley et al. (2002) and associated class proportions for the 13 lithological classes present within the Broken Hill study area. 75
Figure 4.2 Example of T_a spatial distributions for 1, 32 and 1024 clusters 77
Figure 4.3 Mean ranked normalised variable importance for 1, 32 and 1024 $\hat{O}_a$ clusters using all data after the removal of highly correlated variables. 80
Figure 4.4 Comparison of MLA cross-validation accuracies as a function of classification model parameter and number of T_a clusters 81
Figure 4.5 Comparison of MLA mean T_b accuracy with respect to variations in the number of T_a clusters 82
Figure 4.6 Visualisation of the spatial distribution of MLA lithology class predictions using X and Y spatial coordinates (XY Only) as inputs. 85
Figure 4.7 Visualisation of the spatial distribution of MLA lithology class predictions using geophysical data (No XY) as inputs. 86

Figure 4.8 Visualisation of the spatial distribution of MLA lithology class predictions using spatial coordinates and geophysical data (All Data) as inputs 87
Figure 4.9 Comparison of MLA training (using 10-fold cross-validation) and prediction processing times.. 88
Figure 5.1 Schematic diagram of RF decision tree architecture. 99
Figure 5.2 Idealised SVM decision boundary for a 2D (**x**1 and **x**2) non-separable linear binary classification problem. 100
Figure 5.3 Mapped lithologies, after Buckley et al. (2002) within the Broken Hill sample area, western New South Wales, Australia. 102
Figure 5.4 Comparison of the spatial distribution of uncertainty estimated from the class membership probabilities generated by RF and SVM using 0.02, 0.05 and 0.10 T_a sample proportions. 107
Figure 5.5 T_b error-uncertainty thresholds for 0.05 T_a sample proportion 108
Figure 5.6 Comparison of RF and SVM T_b accuracy (100 groups of 1000 samples) as a function of T_a sample size compared with T_b accuracies generated after uncertainty thresholds applied (see Table 5.2) 110
Figure 5.7 Difference between RF and SVM recall and precision rates for the 13 classes in the Broken Hill study area. 112
Figure 5.8 Comparison of the spatial distribution of RF and SVM lithology predictions and unclassified samples identified using uncertainty thresholds using 0.05 T_a sample proportions. 113
Figure 6.1 Map of Tasmania. Black rectangular region indicates the location of the Hellyer–Mt Charter region. 122
Figure 6.2 Typical regional tectonic and geological setting of bimodal-felsic VHMS ore deposits 124
Figure 6.3 Diagram of VHMS a) footwall and b) hangingwall hydrothermal alteration zones in the Hellyer–Mt Charter region 125
Figure 6.4 Interpreted geological map of Hellyer–Mt. Charter region, after Richardson (1994). 127
Figure 6.5 Examples of pre-processed (non-standardised) soil geochemical, airborne geophysical and Landsat ETM+ data used in this study. 132
Figure 6.6 a) Location of the 2100 T_a samples used to optimise RF classification model parameters, select relevant variables and train RF classification models and b) individual class proportions of the total number of samples within the Hellyer–Mt Charter region. 133
Figure 6.7 a) RF variable selection cross-validation accuracy. b) Final list of selected variables in ranked order of relative importance based on mean decrease in RF Gini Index. 134
Figure 6.8 Comparison of: a) interpreted geological map; b) lithology predictions; c) spatial distribution of inconsistencies between the pre-existing interpreted geological map and RF predictions; and d) spatial distribution of RF prediction uncertainty 139
Figure 6.9 Results of SOM unsupervised clustering for the Hellyer Basalt divided into four sub-classes 142
Figure 6.10 SOM sub-classes identified in a) Lower Basalt, b) feldspar-phyric andesite and c) hangingwall andesite units of the QHV. The y-axes of variable frequency plots represent frequency densities. ... 143
Figure 7.1 Study region location and generalised lithologic units, modified from Mineral Resources Tasmania (2011). Table 7.1 provides a summary of class descriptions and abbreviations. 156
Figure 7.2 a) Example of T_a samples locations. Note Qs class was not included in T_a. b) Class proportions within T_b samples, this is approximately equivalent to total area covered by a given class. Note

samples of Qs class not included in T_b for classifier evaluation. Table 7.1 provides a summary of class descriptions and abbreviations. 161
Figure 7.3 Segmented images derived using 100 SOM nodes for a) standard variables and b) texture variables. Note the colour ramp is repeated resulting in duplicated segment colours..... 163
Figure 7.4 Comparisons of spatial-contextual classifier Tb accuracies. 167
Figure 7.5 Selection of the best performing spatial-contextual RF classifiers. Table 7.1 provides a summary of class descriptions and abbreviations..... 168
Figure 7.6 Example of best performing spatial-contextual SVM classifiers. Table 7.1 provides a summary of class descriptions and abbreviations. 169
Figure 7.7 Example of RF and SVM classifications trained on standard variables..... 171
Figure 7.8 Comparisons of mean proportions (across 10 T_a and T_b resamples) of mapped Qs samples classified as classes present within T_a 172
Figure 8.1 RF partial dependence plots showing the relative influence of variables on the prediction presence of QHV units..... 196
Figure A.1 Image (input variable) used for MLA supervised classification training and testing. Random values between 0 and 1 are assigned using a uniform distribution. 255
Figure A.2 Three class example of T_a random sampling (points) and T_b (background) sample structure for trials 1–4. 256
Figure A.3 MLA prediction accuracy distribution boxplots for six classes given “no-information” with four combinations of T_a and T_b class distributions (Table A.2)..... 257
Figure A.4 Trial 2 MLA normalised T_b accuracy density distributions and associated p-values (black lines) for c = 6. T_a class proportions are equal, i.e. 0.1667, T_b class proportions are equivalent to those presented in Table A.3..... 258
Figure B.1 Examples of Variance and Entropy decay curves for different numbers of classes. Note that for c > 8, Variance emphasises higher uncertainties than Entropy. 262

LIST OF ABBREVIATIONS

0D	-	no temporal or spatial (data) dimensions
1D	-	one temporal or spatial (data) dimension
2D	-	two spatial (data) dimensions
3D	-	three spatial (data) dimensions
AEM	-	Airborne Electro-Magnetics
ANN	-	Artificial Neural Networks
ASTER	-	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AVIRIS	-	Airborne Visible/Infra-Red Imaging Spectrometer
BHD	-	Broken Hill Domain
BN	-	Bayesian Networks
BT	-	Boosted Trees
BvSB	-	Best-versus-Second-Best
CSIRO	-	Commonwealth Scientific and Industrial Research Organisation
DEM	-	Digital Elevation Model
DT	-	Decision Trees
ETM	-	Enhanced Thematic Mapper
GIS	-	Geographic Information System
GLCM	-	Grey Level Co-occurrence Matrices
GPR	-	Ground Penetrating Radar
GRS	-	Gamma-Ray Spectrometry
ICT	-	Internet Communication Technology
IEEE	-	Institute of Electrical and Electronics Engineers
kNN	-	*k*-Nearest Neighbours
LDA	-	Linear Discriminant Analysis
LiDAR	-	Light Detection and Ranging
MCMC	-	Markov-chain Monte Carlo
MLA	-	machine learning algorithm
MLC	-	Maximum Likelihood Classifier
MLP	-	Multi-Layer Perceptrons
NB	-	Naïve Bayes

NeCTAR	-	National eResearch Collaboration Tools and Resources
NIR	-	Near Infra-Red
PCA	-	Principal Component Analysis
PKPD	-	Price, Knerr, Personnaz & Dreyfus
PNN	-	Probabilistic Neural Networks
PR	-	post-regularisation
QDA	-	Quadratic Discriminant Analysis
QHV	-	Que–Hellyer Volcanics
RBF	-	(Gaussian) Radial Basis Function
RF	-	Random Forests
RTP	-	Reduced-To-Pole
SAM	-	Spectral Angle Mapper
SAR	-	Single-Aperture Radar
SLP	-	Single-Layer Perceptrons
SOM	-	Self-Organising Maps
SPOT	-	Satellite Pour l'Observation de la Terre
SVM	-	Support Vector Machines
SWIR	-	Short Wave Infra-Red
T	-	labelled samples available for supervised classification
T_a	-	portion of T used for classifier training and validation (training data)
T_b	-	portion of T used for classifier evaluation (test data)
TIR	-	Thermal Infra-Red
TMI	-	Total Magnetic Intensity
UCI	-	University of California, Irvine
VGL	-	Virtual Geophysics Laboratory
VHMS	-	volcanic-hosted massive sulfide
WSG	-	Willyama Supergroup

ACKNOWLEDGEMENTS

This thesis would never have been completed, nor would my sanity have been preserved, without the calm and supportive mentorship of my supervisor Anya Reading. Anya, you taught me many things over the past four years, especially the courage to believe in myself. Special thanks must go to Andrew McNeill for being the secondary supervisor I never really had. Thanks also to Michael Roach for many frantic encounters in the corridor. Sorry I forgot to acknowledge you in my Honours thesis. I hope this makes up for it.

Thanks to Daniel Bombardieri, Jocelyn McPhie, Ron Berry and Jacqueline Halpin for providing comments on my work. Thanks also to the other staff and students, especially Selina Wu, Jeff Steadman and the quiet but resourceful "geofizz" nerds on the top floor (you know who you are), at the ARC Centre of Excellence in Ore Deposits (CODES) and the School of Earth Sciences, University of Tasmania.

Over the past three years I have had the pleasure of visiting and networking with researchers based at other universities across Australia. Thanks to: Malcolm Sambridge, Simone Pilala and Thomas Bodin at the Research School of Earth Sciences, Australian National University; Steve Micklethwaite, Mark Lindsay and Eun-Jung Holden at the Centre for Exploration Targeting, University of Western Australia; and Thomas Landgrebe, Simon Williams and Dietmar Müller from the EarthByte Group, University of Sydney. Meeting you all and discussing different aspects of my research, however briefly, has enriched my Candidature.

I must acknowledge the universe for star dust, gravity and the internet (does that mean I have to thank Telstra?!). I wouldn't be here without you. Also, thanks to those I haven't thanked, especially my caving buddies at STC. I will go caving again one day!

Special thanks to my friends and family. Kyen Knight, for the occasional shoulder to cry on and being there for me when I was a sleep deprived fledgling father. Mum, dad and my brother for feigning interest in my research when really you had no idea what I was talking about. Last but not least, this thesis is dedicated to my beautiful wife Sarah and my cheeky but loving daughter Jasmine. Thanks for believing in me, supporting me and helping me to experience and cherish the most important things in life ... cuddles.

CHAPTER 1 – INTRODUCTION

Recent and rapid technological advances in geoscience data capture and storage are significantly increasing the volume and variety of data available to geoscientists (Kraut & Wettergreen 2010; Bhatia *et al.* 2013; Sellars *et al.* 2013). Vast amounts of continuously collected data, coupled with large volumes of pre-existing data, presents challenges to large scale manual and/or deterministic analysis and interpretation within acceptable time frames (Feyyad 1996; Miller & Han 2001; Kraut & Wettergreen 2010). This is because humans, despite being good at pattern recognition, are limited in the amount of data they can process and analyse at any given time. Furthermore, as models of natural phenomena grow in complexity, i.e. number and variety of inputs (variables) and outputs (target phenomena) increases, understanding the causal relationships amongst model elements becomes increasingly difficult (Reitsma 2010). As a result, we are likely to be missing opportunities to gain knowledge from interactions between multiple layers of disparate data, thus limiting our understanding of complex Earth systems (Feyyad 1996).

Human senses receive information, the source of which is then immediately and subconsciously identified (Ripley 1996). Take for example, the observational and cognitive skills that geologists employ to recognise and assign physical Earth objects and features to predefined categories. The category or class to which an object or feature belongs is determined by a set of consistent rules geologists acquire and learn during years of training. The categorisation of geological features forms the basic premise geologists use to map or model geological phenomena. Geological maps are the fundamental sources of information that geologists use to interpret and understand complex Earth systems.

Our need to understand Earth systems is increasing due to the challenges posed by the burgeoning human population, dwindling resources and climate change. Hence, industry, government and the wider community demand objective, quantifiable and repeatable analyses to aid decision making. Statistical data inference methods such as machine learning provide a uniform, objective standard for the process of data inference, thus, avoiding much of the subjectivity and pitfalls associated with the manual interpretation of data.

1.1. Machine learning

The science of learning from data is a key focus of machine learning. Machine learning combines the fields of statistics and computer science for pattern recognition and data mining applications (Michie *et al.* 1994b; Ripley 1996; Hastie *et al.* 2009). For science based research, pattern recognition is the process of discovering, via automated or semi-automated statistical methods, useful patterns within data (Kotsiantis 2007). Discovered patterns are then used to generate predictions based on similar data (Ripley 1996). The essence of machine-assisted pattern recognition is to provide computers with the ability to adapt their decision structures, based on the characteristics of observed data and generate valid and objective predictions (Schölkopf 2003). Machine learning is an extension of the pattern recognition process. It attempts to provide users with an understanding of the patterns within data (Feyyad 1996; Witten & Frank 2005). Hence, machine learning outputs should be comprehensible in a way that allows interpretations to be formulated in response to the decision structures used to recognise and exploit patterns within data and generate predictions (Feng & Michie 1994; Henery 1994a; Ripley 1996).

Data inference is the act of gaining information, knowledge and ultimately wisdom, from the analysis of raw data using statistical methods (Fotheringham *et al.* 2002; Mucke 2009). The process of data inference can be divided into three levels of understanding (Figure 1.1). The foundation for these levels of understanding is raw data. Successive levels of data inference distil and refine raw data until a complete understanding of the mechanisms controlling the phenomena under investigation is realised (Mucke 2009). The conclusions attained via the process of data inference are subsequently applied to other similar data in

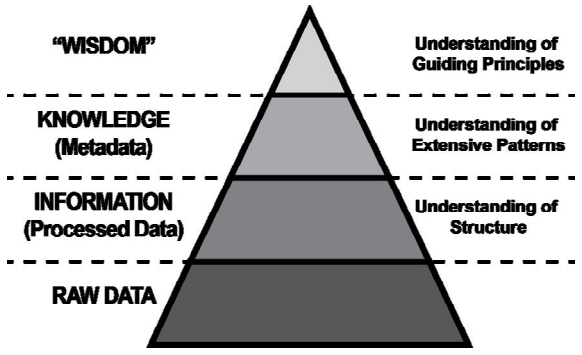

Figure 1.1 Schematic representation of data inference, which is the process of distilling raw data and associated observations into wisdom, or well-informed conclusions, using statistical methods, modified from Mucke (2009).

order to make predictions, formulate interpretations and inform the decision making process (Bousquet *et al.* 2004).

There are two logical approaches to data inference: deductive and inductive reasoning (Feng & Michie 1994; Sivia 1996; Gubbins 2004). Deductive approaches to inference usually involve a researcher (or analyst) who develops a hypothesis (or model) regarding some natural phenomenon. Observational data are then utilised to accept or reject the aforementioned hypothesis (Figure 1.2a). Deductive reasoning can be restrictive in its scope as the results of the analysis are constrained by the pre-formulated hypothesis. In contrast, inductive approaches to inference initially utilise available data/observations to identify patterns. Learned patterns are then employed to develop a range of plausible hypotheses (Figure 1.2b). The advantages of well-performing inductive approaches over deductive approaches to reasoning are based around the ability of computers to rapidly execute repetitive tasks on large digital datasets and infer plausible models without any preconceptions of their form or parameters (Sivia 1996; Burl *et al.* 1998).

Machine learning algorithms (MLAs) are computational tools that provide analysts with a range of statistically sound methods for inductive data inference (MacKay 2003; Bousquet *et al.* 2004). MLAs inductively generate inferences from potentially high-dimensional

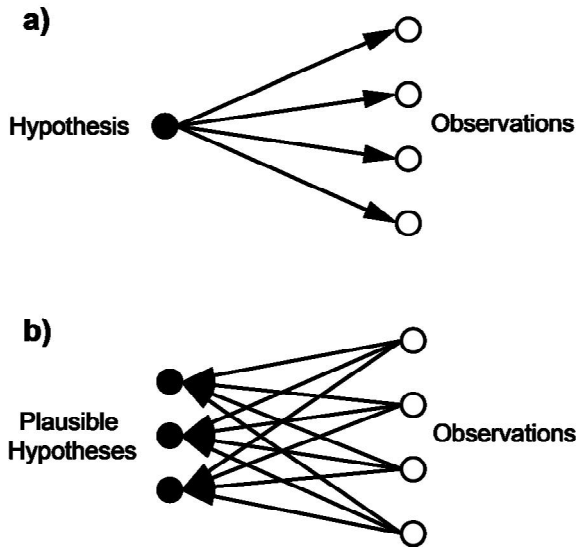

Figure 1.2 Schematic representation of the process of a) deductive reasoning or logic whereby a hypothesis is formulated and tested against data or observations and b) inductive reasoning or logic where multiple plausible hypotheses are formulated by recognising patterns within data or observations, modified from Sivia (1996).

multivariate input data by constructing one or more plausible models that link observed data to the phenomena under investigation (Hastie *et al.* 2009; Chipman *et al.* 2010). Both supervised and unsupervised approaches to machine learning are possible. During supervised learning, a model is induced from training data, or what is known about the problem, that best predicts samples contained within these data. Training is guided by the minimisation of some error or loss function based on the internal architecture of the learning algorithm (Bousquet *et al.* 2004; Hastie *et al.* 2009). In essence, supervised machine learning methods attempting to mimic human reasoning and learning and can be used either for classification or regression problems (Ripley 1996; Burl *et al.* 1998; Witten & Frank 2005; Marsland 2009). In the unsupervised case, a learning algorithm is given unlimited scope in an attempt to find natural groups or clusters in data that best describe its inherent relationships (Ripley 1996; Hastie *et al.* 2009).

1.2. Geological maps

A common task in the geosciences is the process of constructing a map[1] from indirect observations that represents the location and/or properties of geological phenomena (Sambridge *et al.* 2006; Maiti & Tiwari 2010b). This task is traditionally approached using deductive reasoning in which forward modelling and/or inversion techniques are applied. The forward model represents what is known about the problem given a set of physical properties. Inversion then proceeds in an attempt to refine the proposed model to the observed data and the relevant physical properties of the target, i.e. the geological objects or features under investigation (Gubbins 2004).

Geological maps are essentially an abstracted representation of geologically significant features within a 2D spatial reference frame, typically lithological units. Geological maps are a fundamental base layer of information for a wide range of problems and applications (Thomas 2004) such as targeting ore deposits and hydrocarbon resources (Kusky & Ramadan 2002; Jackson 2005; Holden *et al.* 2008; Metelka *et al.* 2011; Shaheen *et al.* 2011), tectonic reconstructions (Dohm *et al.* 2007; Gibson *et al.* 2008; Leverington & Moon 2012), geohazard risk assessment and engineering applications (Tangestani 2004; Ramli *et al.* 2010; Stumpf & Kerle 2011; Sabatakakis *et al.* 2012), geomorphology and hydrology studies (Buselli & Lu 2001; Draskovits & Laszlo 2005) and ecological

[1] In this thesis a geological map is synonymous with a 2D geological model.

modelling (Guisan & Zimmermann 2000; Anderson & Ferree 2010; Brown & Mies 2012). However, geological maps are commonly constructed from the subjective interpretations of a domain expert (the geologist) based on limited field observations. As a result, they are typically imprecise representations of the spatial distribution of geological materials (Bárdossy & Fodor 2001; Grebby *et al.* 2011; Lindsay *et al.* 2013).

The collective experience and knowledge of a geologist is required to relate limited field observations and supplementary data such as geophysical measurements to geological phenomena. The expert subjective knowledge of a geologist, combined with often incomplete observations, the fact that geological and geophysical data contains some form of irreducible or deterministic noise (Scales & Snieder 1998; Ricchetti 2000; Link & Blundell 2003) and the nature of lithological units to exhibit a high degree of interclass similarities and intraclass variability (Ghimire *et al.* 2010; Grebby *et al.* 2011), means that error and uncertainty are an ever present component of geological maps (Bárdossy & Fodor 2001; Gelfort 2006). Despite this, geological maps do not usually provide an indication of the uncertainty associated with its component elements. This is because it is difficult to quantify the confidence with which geologists have used to subjectively interpret the spatial distributions of lithologies during the construction of a geological map (Bárdossy & Fodor 2001; Lindsay *et al.* 2013).

1.3. Research scope and hypothesis

Multivariate geospatial data, collected by airborne or spaceborne remote sensing platforms, complements the use of field observations for the construction of geological maps (Yang *et al.* 1998; Shaheen *et al.* 2011; Leverington & Moon 2012). Given the large quantity of this potentially useful data, MLAs have the potential to provide practical solutions for geological mapping applications in regions where collecting field observations is a costly, time consuming and challenging prospect. Furthermore, as MLAs require minimal user intervention in order to generate outputs, the subjectivity with which predictions are obtained is reduced (Henery 1994a; Ripley 1996). Therefore, MLAs offer an opportunity for geologists to semi-automate the process of creating first-pass geological maps of remote or inaccessible terrain using pre-existing multivariate geospatial data.

Lu & Weng (2007) acknowledge that it is necessary to develop guidelines, for specific fields of research, on the applicability and capabilities of MLAs to achieve the desired

outcomes of research. This is because different learning strategies approach the task of supervised classification in contrasting ways. Given that Kotsiantis (2007) identifies five general machine learning strategies for supervised classification, different MLAs may be better suited to generating geological maps given the spatial context of geoscience data.

Based on the motivations described above and the need to develop an understanding of the capabilities of different MLAs for geological mapping applications, my hypothesis is that it is possible to establish which MLAs are most suited to recurring geological mapping tasks using multivariate geospatial data. Suitability, in this context, is constrained by the requirement of MLAs to generate accurate and plausible map of the spatial distribution of geological features from the integration of disparate multivariate data, while also providing a means of gaining a high level of understanding of complex geological phenomena.

1.3.1. Major research questions to be addressed

I address the following major research questions:

1. What are the advantages and limitations associated with different MLAs in the context of geological mapping applications?

2. What are the best-practice approaches to integrating disparate geoscience data and implementing MLAs for practical real-world geospatial inference problems?

3. Is it possible to assign meaningful levels of confidence to the outputs of MLAs?

4. How does one gain a high level of understanding of the interactions between disparate geoscience data and geological observations using MLAs in order to assist the formulation of meaningful geological interpretations?

With regard to carrying out experiments that answer the research questions posed, it is necessary to develop robust methodologies and streamlined workflows that efficiently implement MLAs and assesses their outputs. For this I have selected the open source statistical programming language, R. R offers users the ability to employ a wide range of existing packages and functions that implement MLAs and assess their outputs via the analysis and visualisation of statistical and spatial data. In addition, R allows users to

develop new functions or modify existing functions for specific purposes, which can be freely distributed to and adapted by other users.

1.4. Thesis structure

This thesis contains two review chapters followed by four chapters, formatted as manuscripts for publication, documenting groups of experiments conducted during the course of my research. The insights arising from the findings detailed in these chapters are synthesised and discussed in terms of their relevance to the hypothesis and research questions outlined previously. This is followed by a summary of key findings and conclusions.

A summary of the structure of this thesis and the contents of chapters are as follows:

- Chapter 2 provides a review of machine learning theory for supervised classification and unsupervised clustering algorithms and best-practice methods for their implementation.
- Chapter 3 reviews previous research using MLAs for geoscience classification applications. Specific focus is placed on the use of spatially distributed geoscience data for geological mapping.
- Chapter 4 – Cracknell & Reading (2014) published in Computers & Geosciences – documents a robust comparison of MLAs, each representing one of the five general machine learning strategies for supervised classification, applied to remote sensing geological mapping. In particular, this chapter focuses on the responses of machine learning strategies to the degree of spatial clustering represented by the training data and the inclusion of explicit spatial information.
- Chapter 5 – Cracknell & Reading (2013) published in Geophysics – addresses the estimation of meaningful measures of MLA prediction uncertainty as a mechanism for identifying ambiguous or erroneous classifications.
- Chapter 6 – Cracknell *et al.* (2014) published in Australian Journal of Earth Sciences – provides a practical example of the integration of spatial geological, geochemical and geophysical data via supervised (Random Forests) and

unsupervised (Self-Organising Maps) MLAs. This work demonstrates the use of machine learning for the critical assessment of pre-existing geological maps and for interpreting complex geological phenomena.

- Chapter 7 – to be submitted for publication in IEEE Transactions on Geoscience and Remote Sensing – investigates methods that provide MLAs with implicit spatial-contextual information for improving geological mapping outcomes.

- Chapter 8 presents a synthesis and discussion of the key findings in light of main research questions and aims. The significance of these findings, with respect to the implementation of MLAs for practical geoscience applications, is presented.

- Chapter 9 is a summation of the key findings and details the conclusions of the research documented in this thesis.

CHAPTER 2 – MACHINE LEARNING THEORY AND IMPLEMENTATION

This chapter documents a comprehensive review of machine learning algorithms (MLAs) for practical data inference problems. Initially, summaries of the motivations and best-practice approaches regarding the three fundamental stages that users must employ in order to implement MLAs for practical supervised classification applications are provided. This is followed by a review of the theory that underpins the five general MLA strategies for supervised learning. The advantages and limitations of these contrasting MLA strategies, as documented in published literature, are considered. This chapter concludes with a summary of MLAs for unsupervised learning and includes a description of the theory behind several well-known and popular clustering algorithms.

2.1. Machine learning

Machine learning exploits the advantages of committing computers to the task of learning by utilising their abilities to rapidly execute calculations, store information and receive instruction (Sivia 1996; Burl *et al.* 1998). MLAs attempt to recognise patterns in data via an automatic adaptive approach and then apply the learned relationships to other similar data. MLAs are able to inductively generate predictions for classification and regression problems and are especially useful where the process under investigation is represented by high-dimensional multivariate input data (Witten & Frank 2005; Kotsiantis 2007; Kanevski *et al.* 2009). Predictions generated by MLAs should be, as a bare minimum, consistent and comparable to the abilities of human analysts (Burl *et al.* 1998). Hence, MLAs must be able to handle a wide variety of problems and, given enough information, be extremely accurate, reliable, repeatable and efficient when applied to real-world data (Henery 1994a; Michie *et al.* 1994b). The ultimate goal of machine learning is to support humans in the process of gaining insight into the complex relationships between natural phenomena such that interpretations can be formulated in response to the induced decision structures. In other words, the outputs of MLAs should facilitate human understanding and assist in the development of well-informed and robust decisions (Feng & Michie 1994; Henery 1994a; Ripley 1996).

MLAs generate predictions by linking input data (variables) to desired outcomes (targets, Friedman 1997; Chipman *et al.* 2010). The nature of the inference problem is defined by the target data types provided to the MLAs during training. For example, the aim of classification is to generate predictions for a set of discrete classes that represent unordered descriptive categories. In contrast, the aim of regression is to predict ordered numeric values (Marsland 2009). This chapter does not address machine learning for regression problems as the objective of my research is to investigate MLAs for the classification of discrete categories representing geological features such as lithologies.

2.1.1. Supervised versus unsupervised learning

MLAs are employed to generate inferences from data (Friedman 1997) using either supervised or unsupervised approaches to learning. Supervised learning utilises training data and can be defined as function approximation (Kotsiantis 2007; Hastie *et al.* 2009) where the aim is to minimise an error or loss criterion (Kuncheva 2004; Marsland 2009). Training data comprises a set of discrete and unordered labels indicating known outcomes or observations. These data are used to induce or train a classification model that links variables to the classes present in the training data while minimising the error criterion. Once constructed, a supervised classification model enables one to predict the class labels of samples previously unseen by the MLA (Kuncheva 2004; Hastie *et al.* 2009). In contrast to supervised approaches to learning, unsupervised learning only has a set of variables and no observed or prior knowledge of desired outcomes, i.e. it is data-driven (Ripley 1996). The lack of prior knowledge with which to assess the results of unsupervised learning means that the quantitative assessment of outputs is infeasible. Thus, the aim of unsupervised learning is not to minimise an empirical error function, it is to identify natural, coherent groups within data as a means of gaining insight into how these data are organised (Kuncheva 2004; Hastie *et al.* 2009).

2.2. Supervised classification

Supervised classification can be formally described as linking one domain (input data) to another (target classes) via a discrimination function (without considering noise or random error):

$$y = f(\mathbf{x}). \qquad [2.1]$$

Inputs (variables) are represented as d vectors of the form $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d \rangle$ and y is a finite set of c class labels $\{y_1, y_2, \ldots, y_c\}$ as indicated by the labelled data, *T*. Given instances of $\mathbf{x}$ and y, supervised classification attempts to induce or train a classification model f', where $f' \quad f$ such that,

$$\hat{y} = f'(\mathbf{x}), \qquad [2.2]$$

which links variables to classes (Gahegan 2000; Hastie *et al.* 2009; Kovacevic *et al.* 2009). The form of f' is induced by some internal strategy that minimises the empirical error (Kohavi 1995; Schölkopf 2003; Hastie *et al.* 2009). In other words, given the architecture of a particular MLA, decision structures are induced that reduce the cost of classification, i.e. misclassifications, to a minimum given the information within *T*. In practice, as we only have class labels for a limited set of data, it is necessary to divide available data into independent groups for MLA training, $\hat{O}_a$ (training data) and evaluation, $\hat{O}_b$ (test data, Witten & Frank 2005; Hastie *et al.* 2009).

2.2.1. Classification strategies

This section outlines fundamental theory behind the five general machine learning strategies for supervised classification (Kotsiantis 2007): statistical learning algorithms; instance-based learners; logic-based learners; Support Vector Machines; and Perceptrons.

2.2.1.1. Statistical learning algorithms

Statistical learning algorithms construct an explicit statistical model that estimates the probability that a sample belongs to a particular class (Domingos & Pazzani 1997; Guyon 2009). These include, Linear (and Quadratic) Discriminant Analysis (LDA, QDA) which use density estimation techniques to find linear (or quadratic) combinations of variables that best separate multiple classes (Kotsiantis 2007). Alternatively, Bayesian models such as Naïve Bayes (NB) and Bayesian Networks (BN) provide methods for probabilistic supervised classification. These methods are superior to LDA or QDA for problems represented by high-dimensional inputs variables that exhibit non-linear relationships between multiclass targets (Witten & Frank 2005).

NB is a well-known probability density estimation technique for classification problems. It is recommended as a base-level classifier for comparison with other algorithms (Henery 1994a; Domingos & Pazzani 1997; Guyon 2009; Hastie *et al.* 2009). NB employs Bayes

Theorem, which states that the probability, P, that a class y is correct given the data **x** using,

$$\mathrm{P}(y|\mathbf{x}) = \frac{\mathrm{P}(\mathbf{x}|y) \times \mathrm{P}(y)}{\mathrm{P}(\mathbf{x})}. \qquad [2.3]$$

$\mathrm{P}(y)$ is the prior probability and represents our knowledge of the problem, i.e. T_a. $\mathrm{P}(\mathbf{x}|y)$ is the probability that the data is true given the class and represents the likelihood function used to update the prior via empirical analysis. $\mathrm{P}(\mathbf{x})$ indicates the data used during analysis, which is constant for all classes, thus it can be ignored for most applications (John & Langley 1995; Sivia 1996; Domingos & Pazzani 1997). To simplify the problem, NB assumes that for a given class the input variables are independent of each other. This assumption yields a discrimination function indicated by the products of the joint probabilities that the classes are true given the data for sample *i*, represented by a vector of inputs **x** with values u_i:

$$f\ (\mathbf{x}_i) = \mathrm{P}(\hat{y}) \prod_{d=1}^{v} \mathrm{P}(\mathbf{x} = u_i|\hat{y}). \qquad [2.4]$$

Using this logic, the problem can be reduced to finding a vector of probabilities where the Bayes optimal classifier is represented by the maximal class probability (Figure 2.1, Duda & Hart 1973; Molina *et al.* 1994; Hastie *et al.* 2009).

NB is not prone to overfitting (see Section 2.2.2.3, Guyon 2009) and often outperforms more sophisticated alternatives unless **x** contains correlated and hence non-independent

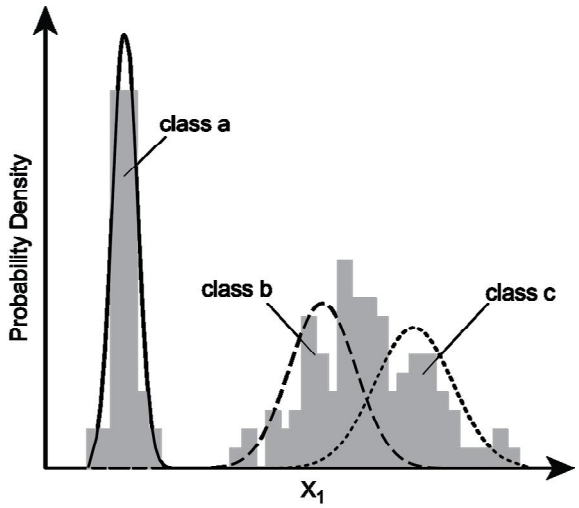

Figure 2.1 An example of NB estimated joint (normal) class conditional probability densities for one discretised continuous variable and three classes.

variables (Domingos & Pazzani 1997; Witten & Frank 2005; Tan *et al.* 2006; Hastie *et al.* 2009). Non-independent variables skew the NB learning process such that discrimination will effectively be concentrated on the set of dependent variables (Witten & Frank 2005). Furthermore, NB is limited where a given variable does not occur in conjunction with every class in T_a. In this case, the probability of a variable leading to the class in question would be zero. As the final probabilities for a given class are the product of all the probabilities, the resultant will also be zero (Witten & Frank 2005; Kotsiantis 2007). The default NB formulation assumes normally distributed continuous variables. In cases where this assumption does not hold, kernel density estimation procedures that do not assume a particular distribution will improve performance (John & Langley 1995; Witten & Frank 2005).

BN (or Directed Acyclic Graphs, Pearl 1988; Ripley 1996) are an extension of NB that does not assume independent variables (Witten & Frank 2005). BN construct a graphical (network) model where nodes represent a variable and the links between nodes represent the correlation between variables. The local conditional probability distributions of a variable given its parents are represented by a set of tables, one for every linked variable. From these tables, joint probability distributions are calculated based on the independence structure within the network (Friedman *et al.* 1997). In circumventing the assumption of independence, calculating posterior probabilities using all data becomes computationally expensive or infeasible for large numbers of variables (Kotsiantis 2007).

An alternative approach to inducing BN is to approximate posterior probabilities using Markov-Chain Monte Carlo (MCMC) methods. MCMC, in the context of BN, is used to randomly sample variables and then weight these samples by their likelihood functions. As the number of samples increases the closer the estimation converges on their expected values (Marsland 2009). Further reductions in processing time can be obtained by only assessing a variable, its child nodes and the parent nodes of those child nodes, termed a Markov blanket. The nodes within a Markov blanket are conditionally independent of all other nodes, therefore, the variable they represent is irrelevant and not required to estimate the likelihood function (Witten & Frank 2005; Marsland 2009).

Despite methods available to reduce the computational cost of implementing BN, this algorithm is still relatively inefficient with respect to NB (Kotsiantis 2007; Marsland 2009). Furthermore, rigorous empirical experiments indicate that there is often no

significant difference between the predictions generated by NB or those of BN (Dumais *et al.* 1998). Experiments conducted by Friedman *et al.* (1997) showed that in many cases NB generated significantly more accurate predictions than BN.

2.2.1.2. Instance-based learners

Instance-based learning algorithms, also known as lazy learners, feature three unique characteristics that distinguish them from other learning strategies: (1) T_a is stored until required to generate predictions for individual samples; (2) predictions are formulated by combining samples in T_a; and (3) they discard predictions once they have been generated (Aha 1997; Wettschereck *et al.* 1997). Unlike many other MLAs, which construct a global classification model based on all the samples in T_a, instance-based learners train a single classifier for every sample requiring prediction. The most commonly used instance-based learning algorithm is *k*-Nearest Neighbours (kNN, Atkeson *et al.* 1997).

For an individual sample that requires classification, the kNN algorithm queries a selected representative, i.e. the closest neighbours, of T_a that reside within a local region of variable space (Bottou & Vapnik 1992; Atkeson *et al.* 1997). This algorithm, developed by Fix & Hodges (1951) and Cover & Hart (1967) is based on the assumption that the sample to be classified is most likely to be proximal to the most abundant class contained within neighbouring observations in *d*-dimensional variable space (Henery 1994a; Wettschereck *et al.* 1997; Witten & Frank 2005; Tan *et al.* 2006). Thus, the majority class within neighbouring T_a samples defines the predicted class label (Figure 2.2). In situations where multiple maximal classes are identified within a local neighbourhood one of these classes

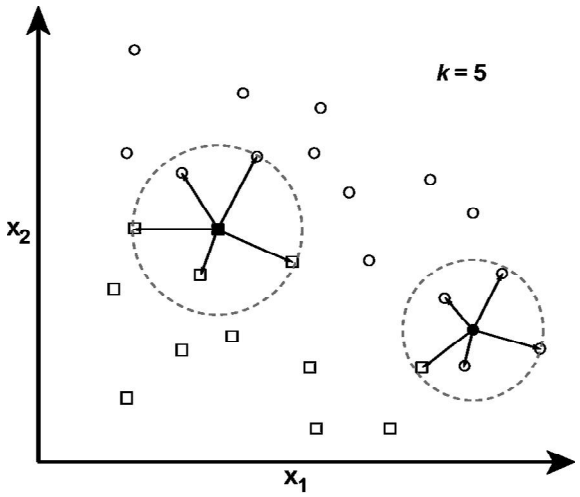

Figure 2.2 Schematic diagram of kNN classifications for a binary classification task in 2D variable space. Filled symbols represent predicted class for an unlabelled sample.

is randomly selected (Kotsiantis 2007).

The kNN algorithm can be defined formally as:

$$\hat{y} = \max_c\left(\sum_{j=1}^{k} I(y_j = c)\right), \qquad [2.5]$$

where $I(y_j = c)$ is an indicator function that identifies the *k* neighbouring T_a samples equal to class *c*. It is important to select an appropriate value of *k*, the number of nearest neighbours to assess. Too small a value of *k* often results in underfitting (see Section 2.2.2.3) and susceptibility to noise. Conversely, too large a value of *k* will generate classifiers that are over-fitted and more likely to misclassify a T_b sample due to the presence of samples belonging to another class. Typically, methods such as cross-validation are required to select appropriate values of *k* (Tan *et al.* 2006; Kotsiantis 2007; Hastie *et al.* 2009).

In conjunction with the selection of an appropriate value of *k*, a similarity or distance measure is used to establish the *k* closest neighbours in variable space. The most commonly used similarity measure is the Euclidian distance metric, although other metrics that measure proximity between samples in variable space are possible. These metrics include the Manhattan and Minkowsky distances (Table 2.1, Hechenbichler & Schliep 2004; Kotsiantis 2007). As the proximity of T_a samples is established using relative distances in variable space scaling can be an issue. Therefore, it is recommended that continuous variables be standardised to a common scale, especially if the variables used as input are measured in different units (Molina *et al.* 1994; Kotsiantis 2007; Hastie *et al.*

Table 2.1 Common distance metrics used to measure the separation distance between samples in multi-dimensional variable space, after Hechenbichler & Schliep (2004) and Kotsiantis (2007).

Euclidian	$D(x,y) = \sqrt{\sum_{i=1}^{n} \lvert x_i \quad y_i \rvert}$
Manhattan	$D(x,y) = \sum_{i=1}^{n} \lvert x_i \quad y_i \rvert^{2}$
Minkowsky	$D(x,y) = \sqrt[r]{\sum_{i=1}^{n} \lvert x_i \quad y_i \rvert^{r}}$

2009). In addition, it is easier to implement kNN using numeric variables as the distance between values is explicit rather than with categorical variables where some notion of similarity, i.e. normalisation, is required to provide a measure of scale (Witten & Frank 2005).

As all variables are treated equally by kNN, it is affected by redundant, irrelevant or noisy variables (Henery 1994a; Wettschereck *et al.* 1997; Hechenbichler & Schliep 2004). In addition, the computational cost of finding neighbours and storing T_a can be excessive for large numbers of samples (Molina *et al.* 1994; Hastie *et al.* 2009). Despite this, kNN has been shown to perform well in situations where decision structures, i.e. the separation of classes in variable space, are highly irregular (Hastie *et al.* 2009). Furthermore, unlike logic-based learners or Perceptrons, kNN will usually generate stable predictions in light of changes in T_a (Breiman 1996).

The standard kNN algorithm assumes that the k neighbouring T_a samples to the sample requiring prediction are of equal importance. Equal weighting of k proximal samples can incorporate redundant, irrelevant or noisy variables in the resulting classification model resulting in poor performance (Wettschereck *et al.* 1997). Moreover, the standard kNN algorithm is vulnerable to high bias (see Section 2.2.2.3) when faced with high-dimensional inputs (Friedman 1994; Hastie & Tibshirani 1996). Adjusting training sample weights using monotonically decreasing distance decay functions emphasises the contribution of proximal neighbours on predictions, thus dampening the potential for over or underfitting (Hechenbichler & Schliep 2004). Sample weighted kNN proceeds using a weighted majority vote:

$$\hat{y} = \max_c \left(\sum_{j=1}^{k} w_j I(y_j = c) \right), \qquad [2.6]$$

where w_j indicates a weighting function based on the distance metric selected.

Adaptive kNN algorithms utilise methods that identify different values of k for a given class and/or sample. Alternatively, they modify the geometry of the search neighbourhood based on input variability (Hastie *et al.* 2009). Dietterich & Wettschereck (1994) outline several approaches to computing optimal local values of k. The standard local adaptive kNN (called $kNN_{unrestricted}$), on which all other variants are based, uses cross-validation to identify k values that correctly predict individual T_a samples. To classify a given sample, the M nearest T_a samples are identified and the k neighbours that correctly classify the

majority of these samples are selected. Friedman (1994) describes the Flexible Metric Nearest Neighbour algorithm, an adaptive form of kNN that adjusts the geometry of the search region based on the local relevance of variables via recursive partitioning. An optimal *k* value is selected by iteratively adjusting distance metrics orthogonal to variable space axes. This generates elongate search neighbourhoods in directions where there is very little observed variability in class probabilities (Molina *et al.* 1994; Hastie *et al.* 2009). Using a similar approach, Hastie & Tibshirani (1996) developed a Discriminant Adaptive-nearest Neighbour algorithm that incorporates local LDA to estimate optimal distance metrics for adjusting search neighbourhoods parallel to the axes of variable space.

2.2.1.3. Logic-based learners

Logic-based learners use a set of criteria or rules in order to generate predictions. The simplest forms of logic-based learners incorporate a series of Boolean decisions ("and", "or", "if else") that, in the context of machine learning, are automatically generated and define classification rules (Feng & Michie 1994; Witten & Frank 2005). Logic-based classification models map classes to variables by dividing T_a into discrete partitions of self-similarity (Henery 1994a). These partitions must classify a minimum number of T_a samples belonging to the classes split under the current rule (Feng & Michie 1994). Decision Trees (DT) provide a concise representation of logic-based learners (Witten &

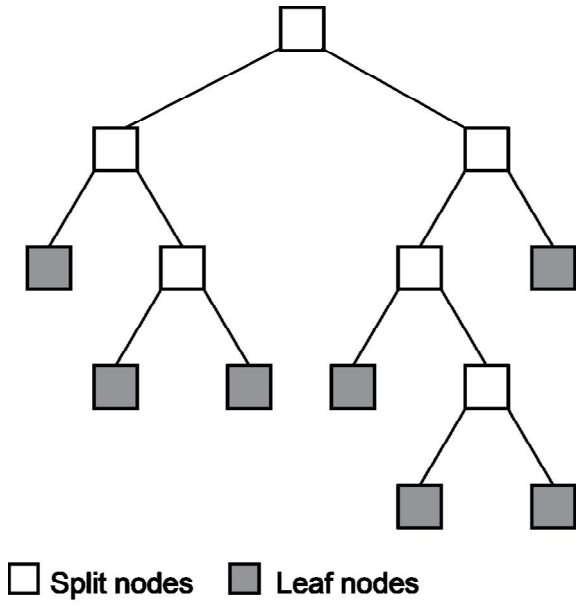

Figure 2.3 Schematic representation of binary DT classifier architecture. Split nodes partition inputs based on some threshold (for continuous data). Leaf nodes terminate the partitioning process and generate a class prediction.

Frank 2005) and are the foundation for more sophisticated algorithms that use this learning strategy. Breiman *et al.* (1984) pioneered the widespread use and implementation of DT under the pretence that they generated accurate predictions while also providing a basis for understanding the structural relationships between target classes and input variables.

DT apply conditions to a particular variable and then splits T_a based on thresholds or Boolean decision rules (Feng & Michie 1994; Witten & Frank 2005). DT comprise a series of nodes in a cascading top-down inductive network. From each "parent" node data is split into "child" nodes that are either branching nodes, where data is subjected to splitting again; or leaf nodes, where the resulting class is put forward as a prediction (Figure 2.3). DT are constructed from a root node, the initial starting node of the trees and grown recursively unlit some stopping criterion is reached, such as the minimum purity of the leaf nodes (Breiman *et al.* 1984). Numeric inputs are dealt with by spitting on threshold values, i.e. using greater than or less than rules, at an established point within the numeric range of a given variable (Witten & Frank 2005).

A special form of DT are the Classification and Regression Trees (CART) developed by Breiman *et al.* (1984). CART are binary DT that split parent nodes into two child nodes, thus avoiding the rapid fragmentation of data as can occur with multi-way spits (Hastie *et al.* 2009). CART use the information gain ratio (Gini Index) to defined an optimum splitting threshold for branching nodes (Breiman *et al.* 1984). The Gini Index calculates the information purity of child nodes with respect to that of their parent node by comparing the ratio of classes passed onto a child node. The Gini Index is defined as:

$$Gini(t) = \sum_{c=1}^{j} g_c\,(1 \quad g_c), \qquad [2.7]$$

where g_c is the probability or the relative frequency of class *c* at node *j* and is given by:

$$g_c = \frac{n_c}{n}, \qquad [2.8]$$

where n_c is the number of samples belonging to class *c* and *n* is the total number of samples within a particular node. For each candidate split, the threshold t that defines maximum reduction in class heterogeneity is selected (Breiman *et al.* 1984; Waske *et al.* 2009; Waske *et al.* 2012). Hence, if there is only one class at a given node the purity is at a maximum and the information within the node is at a minimum. In other words, there is no

possible benefit (information) for splitting a node with only one class (Witten & Frank 2005).

CART employ a tree pruning method that strives to generate small trees while maintaining accurate estimates of the true probabilities of misclassification. Small trees are advantageous as they are often easier to interpret, reduce computational cost and are less prone to overfitting (Kotsiantis 2007). Pruning trees by replacing a sub-tree (series of nodes linked to a common parent node) is designed to address the trade-off between simplicity and accuracy (Feng & Michie 1994; Witten & Frank 2005).

DT are challenged by the presence of missing values (Witten & Frank 2005) and sensitive to small changes in T_a. In these situations, DT can generate significantly different splitting thresholds, which result in high variance (see Section 2.2.2.3). Ensemble methods, or a committee of classifiers, minimise the instability of DT by combining the results of multiple classifiers (Kuncheva 2004; Hastie *et al.* 2009). There are two basic algorithms that use multiple DT to cast a vote on the predicted class: Random Forests™ (RF), trademark of Leo Breiman and Adele Cutler; and Boosted Trees (BT, Banfield *et al.* 2007; Guyon 2009; Waske & Braun 2009).

RF, developed by Breiman (2001), is an ensemble classification scheme that utilises a majority vote for class association based on the results of multiple randomised DT, known as a forest. Randomness is introduced into the algorithm by randomly subsetting a predefined number of input variables (*mtry*) to split at each node of individual DT and by bagging (bootstrap aggregation). Bagging (Breiman 1996) generates T_a samples for each tree by sampling with replacement a number of samples equal to the number of instances in T_a. This equates to approximately two-thirds of samples available for training while the remaining samples are used for evaluation. Bagging is reported to improve classification predictions as long as they are not stable in the presence of altered T_a (Breiman 1996). The Gini Index is used by RF to determine a "best-split" threshold at each node of individual DT. RF grows multiple DT and is generally insensitive to noise and model overfitting (Breiman 2001; Tan *et al.* 2006).

The basic premise of BT can be found in the AdaBoost algorithm (Freund & Schapire 1997). In this case, combinations of "weak" classifiers, i.e. classifiers whose error rate is marginally better than a random guess, are trained using modified weights that are iteratively adjusted. Adjustments to T_a weights are designed to place more emphasis on

samples that are difficult to classify by assigning higher weights to samples misclassified in the previous iteration, thus forcing BT to concentrate on correctly classifying these samples. The final result is obtained by combining all iterations (trained DT) using a weighted majority vote that emphasises the influence of more accurate DT in the ensemble (Freund & Schapire 1996; Banfield *et al.* 2007; Hastie *et al.* 2009).

2.2.1.4. Support Vector Machines

Support Vector Machines (SVM) are popular MLAs first described by Vapnik (1995; 1998). They have the ability to define decision boundaries between classes in a high-dimensional variable space by maximising the margin between two classes (Karatzoglou *et al.* 2006; Hsu *et al.* 2010). This approach to separating classes has been shown to reduce an upper bound on prediction error (Kotsiantis 2007). Furthermore, SVM training will always find a global minimum in contrast to other algorithms such as Artificial Neural Networks (Burges 1998; Kotsiantis 2007).

Basic SVM theory states that for a linearly separable dataset containing points from two classes there are an infinite number of hyperplanes that divide these classes. The hyperplane, *h*, is formally defined by:

$$\boldsymbol{w}\ \ \mathbf{x} + b = 0, \qquad [2.9]$$

where $\boldsymbol{w}$ and *b* are solved via quadratic optimisation. The separation of two classes via *h* is achieved using only a subset of T_a instances known as support vectors (Tan *et al.* 2006). SVM model complexity is unaffected by the number of variables because the number of support vectors identified is usually small. Thus, SVM circumvents the "curse-of-dimensionality" (see Section 2.2.3.1), i.e. the proliferation of variables causing intractable complexity and overfitting (Burges 1998). For this reason, SVM are well suited to deal with learning tasks where the number of variables is large with respect to the number of samples in T_a (Kotsiantis 2007).

The maximum margin *M* (distance), equal to $\frac{\mathbf{2}}{\|w\|}$ perpendicular to *h* that separates the classes in question is taken to represent the optimal decision boundary. Obtaining *M* is equivalent to minimising the objective function:

$$min_{\boldsymbol{w},b}\ \frac{\|\boldsymbol{w}\|^2}{2}, \qquad [2.10]$$

$$\text{subject to } y_i(\boldsymbol{w}.\mathbf{x}_i + b) \quad 1\ , i = 1,2,\dots,c.$$

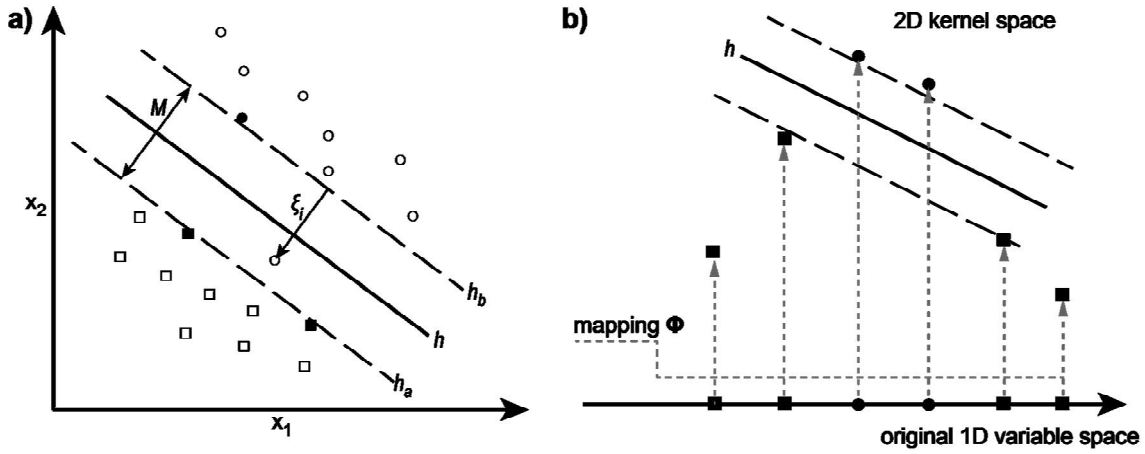

Figure 2.4 SVM schematic diagrams of a) separating hyperplane in 2D variable space, where indicates class a and class b. Filled symbols represent support vectors. b) Kernel transformation example from 1D variable space to 2D kernel space, modified from Smirnoff et al., (2008), Kovacevic et al. (2010) and Yu et al. (2012).

Classifications are obtained by assessing which side of *h* the sample requiring a class label resides (Burges 1998).

In non-separable linear cases, SVM find *h* while incorporating a cost parameter *C*, which adjusts the penalty associated with misclassifying support vectors (Figure 2.4a). High values of *C* generate more complex prediction functions in order to misclassify as few support vectors as possible by way of a high penalty on error (Burges 1998; Karatzoglou *et al.* 2006). The objective function must be modified to incorporate this penalty term for wide margined decision boundaries with misclassified T_a:

$$min_{\boldsymbol{w},b,\xi} \frac{\|\boldsymbol{w}\|^2}{2} + C \sum_{i=1}^{N} \xi_i, \qquad [2.11]$$

$$\text{subject to } y_i(\boldsymbol{w}.\mathbf{x_i} + b) \quad 1 \quad \xi_i, i = 1,2,\dots,c,$$

where slack variables ξ_i 0 represent the distance to misclassified support vectors from their respective marginal hyperplanes (Hsu *et al.* 2010).

For non-linear cases, SVM use an implicit transformation of input variables via a kernel function *kern*:

$$kern(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle, \qquad [2.12]$$

which returns the inner product between the positions of pairwise compared input variables ($\mathbf{x}_i$ and $\mathbf{x}_j$) in variable space (Figure 2.4b). The kernel function allows SVM to handle non-

Table 2.2 Common kernel functions for SVM, after Karatzoglou et al. (2006).

Linear	$kern(\mathbf{x}, \mathbf{x}) = (\mathbf{x}, \mathbf{x})$
RBF	$kern(\mathbf{x}, \mathbf{x}) = \exp(\ \sigma \lvert \mathbf{x}, \mathbf{x} \rvert^2)$

linear relationships efficiently between classes and variables by projecting samples from the original *d*-dimensional variable space into a potentially infinite dimensional kernel space (Burges 1998; Hsu *et al.* 2010). It is important to select an appropriate kernel function from which T_a will be classified (Burges 1998; Kotsiantis 2007). Linear and Gaussian Radial Basis Function (RBF) kernels (Table 2.2) offer good first-choice kernels for most applications (Karatzoglou *et al.* 2006).

The architecture of SVM described above deals with binary classification tasks. SVM can be extended to multiclass problems by combining multiple classifiers (Kotsiantis 2007). Two methods for combining SVM classifiers appear in the literature, both of which are based on a majority vote. The one-against-all method trains a number of classifiers equal to the number of classes in T_a, with each classifier separating one class from the rest (Burges 1998). In contrast, the one-against-one method constructs $\frac{c\,(c-1)}{2}$ pairwise classification models, separating one class against another. Despite constructing more classifiers, the one-against-one method has been shown to efficiently generate robust classifications (Hsu & Lin 2002; Karatzoglou *et al.* 2006; Kovacevic *et al.* 2010).

Unlike other machine learning strategies, SVM do not employ density estimation to discriminate classes. In contrast, SVM exploit the geometrical characteristics of data by assessing only support vectors in order to geometrically define decision boundaries (Burges 1998; Melgani & Bruzzone 2004). However, SVM performance is sensitive to the choice of kernel, the size of the kernel, ó, used construct the transformed variable space and *C* (Hsu *et al.* 2010). Therefore, Hsu *et al.* (2010) suggest using *k*-fold cross-validation to assist in establishing optimal SVM parameters for the intended application. In addition, it is recommended that input variables be standardised to a common scale (Hsu *et al.* 2010).

2.2.1.5. Perceptrons

Perceptron MLAs abstractly model the ability of biological nervous systems to recognise patterns and objects. Perceptrons include a suite of algorithms known as Artificial Neural Networks (ANN, Rojas 1996; Hastie *et al.* 2009). ANN are composed of a network of primitive functions (artificial neurons), which were first described by McCulloch and Pitts (1943). These so called McCulloch–Pitts neurons (Figure 2.5a) consist of two components, a weighted sum of its inputs followed by an activation function, which can be generalised to the form (Rohwer *et al.* 1994; MacKay 2003):

$$y_i = f_k\left(\sum_i \boldsymbol{w}_{ji}\mathbf{x}_i\right), \qquad [2.13]$$

where $\boldsymbol{w}_{ji}$ is the adjustable weight for the i^{th} instance and $\mathbf{x}_i$ indicates one of the input variables. The "activation" function f_k can be any non-linear function, e.g. step or sigmoidal (Rohwer *et al.* 1994; Rojas 1996), capable of receiving multiple weighted inputs. Function inputs are evaluated in terms of their success at discriminating the classes in T_a. This is achieved via an adjustable threshold u_k where if the sum of the weighted inputs is above a given threshold the output is 1, else it is 0 (Rohwer *et al.* 1994; MacKay 2003; Witten & Frank 2005; Kotsiantis 2007).

Single-Layer Perceptrons (SLP) are networks of McCulloch–Pitts neurons that linearly segment variable space into regions for classification (Rohwer *et al.* 1994; MacKay 2003). SLP use the Perceptron Learning Rule proposed by Rosenblatt (1962) to establish multiclass decision boundaries by iteratively learning suitable weights for the neurons within the network. Weights are adjusted such that errors are minimised via an objective function that measures how well the weighted network predicts the correct class for a given sample in T_a. The learned network weights are then applied to new data in order to generate predictions (MacKay 2003). If the division of input space incurs an error, network weights are adjusted to arrive at a perfect solution, if and only if, the data is linearly separable (Rohwer *et al.* 1994; Witten & Frank 2005). However, many real-world problems cannot be represented by this simple linear system. Therefore, the estimation of non-linear decision structures is achieved by combining one or more hidden layers in a network, so called Multi-Layer Perceptrons (MLP, Rohwer *et al.* 1994).

MLP are routinely implemented non-linear ANN (Kotsiantis 2007; Hastie *et al.* 2009). They consist of weighted connections between a layer of input neurons, one or more layers

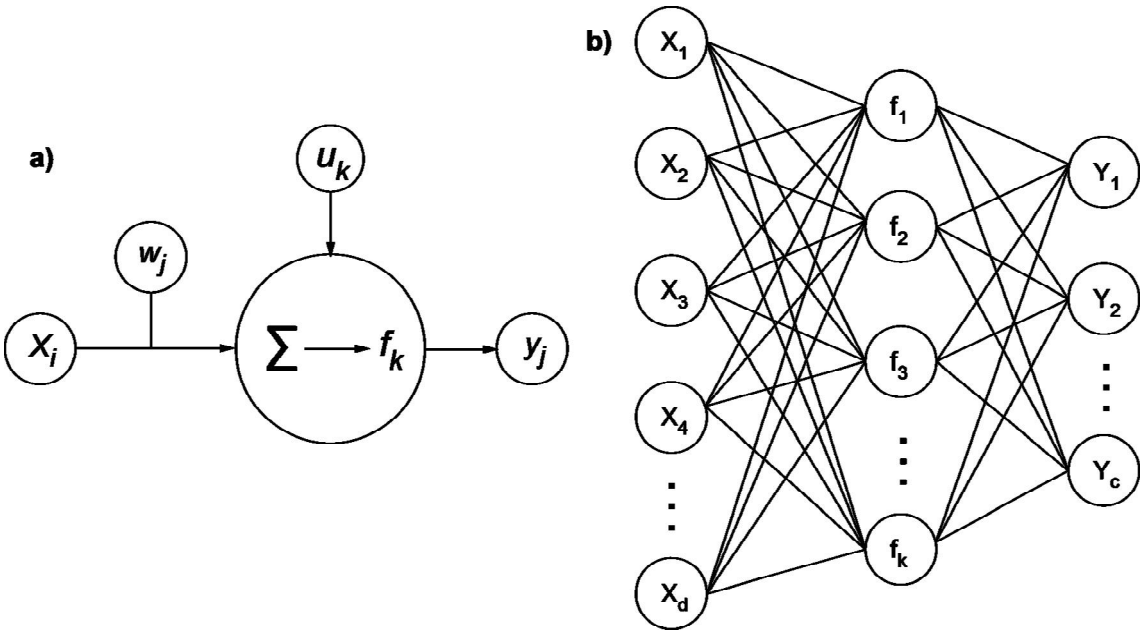

Figure 2.5 Schematic diagram of a) single McCulloch–Pitts neuron, modified from Rohwer et al. (1994) and b) single hidden-layer feed-forward network (MLP), modified from Hastie et al. (2009).

of hidden neurons and one output layer of neurons. The simplest form of MLP are feed-forward networks with a single hidden layer (Figure 2.5b). They are feed-forward in the sense that signals are allowed to travel only from input to output layers and that y are a function of $\mathbf{x}$ (Rohwer *et al.* 1994; MacKay 2003; Kotsiantis 2007). The hidden middle layers of networks are not visible to either the inputs or output layers, forcing the network to make a simple model of the system under investigation (Rohwer *et al.* 1994). In multiclass classification problems, target classes can be represented by a vector where the correct class is assigned 1 and all others 0. In these cases, outputs are coupled such that they sum to 1 with the result interpreted as class membership probabilities (MacKay 2003). MLP routinely utilise back-propagation (also known as the Delta Rule) to facilitate the minimisation of training error and converge on a solution (Hastie *et al.* 2009). Output neuron training error is propagated back through a linearised version of the network, which is used to establish the error gradient. Typically, convergence proceeds until the error of back-propagation reaches a decay threshold (Rohwer *et al.* 1994; Rojas 1996; Kotsiantis 2007).

In order to optimise MLP, i.e. feed-forward back-propagation networks with a single hidden layer, there are two model parameters that must be tuned and selected: (1) the number nodes in a hidden layer; and (2) a parameter that sets the error gradient decay threshold. Increasing the number of hidden nodes will result in arbitrarily complex

decision boundaries which can lead to overfitting or an inability to converge on a result. Too few hidden nodes make the network prone to underfitting and may reduce the ability of MLP to capture non-linearity in data (Rohwer *et al.* 1994; Kotsiantis 2007). The decay threshold parameter indirectly regularises the search for a global error minimum, preventing the construction of over-fitted classification models. Furthermore, it is recommended that input variables are standardised because the scale of the input variables directly affects the scaling of weights in the output layer of the network (Hastie *et al.* 2009).

2.2.2. Supervised classifier implementation

After available data (input variables and target classes) have been collated there are three key stages for the implementation of MLAs to supervised classification problems: (1) data pre-processing; (2) classifier training; and (3) prediction evaluation (Figure 2.6). All three stages require careful preparation and execution in order to induce optimally performing classification models and to provide a robust assessment of their expected performance on new data (Feng & Michie 1994; Kuncheva 2004).

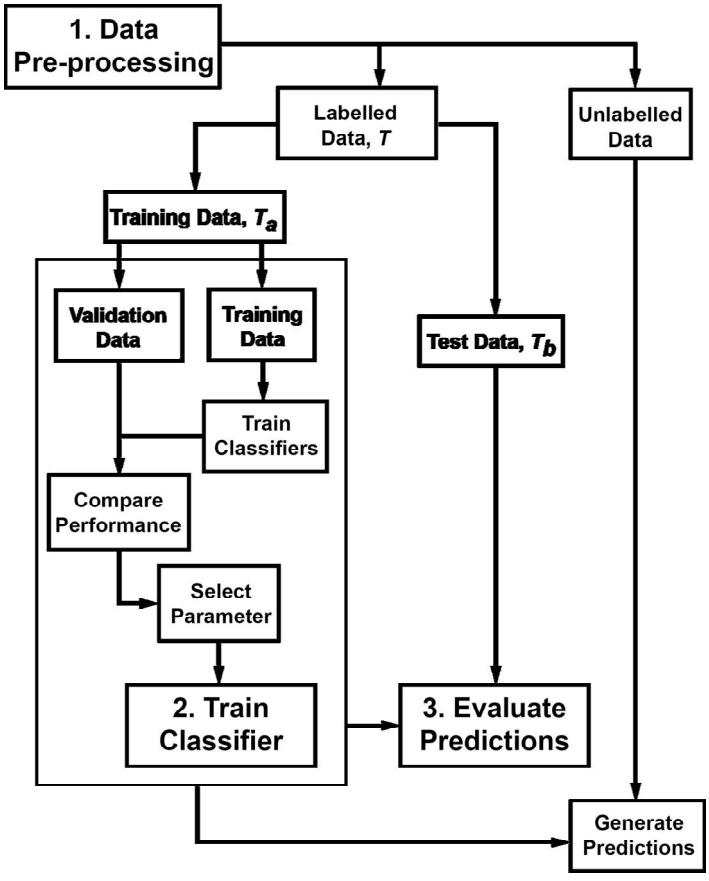

Figure 2.6 Generalised workflow for machine learning supervised classification.

2.2.2.1. Data pre-processing

Data pre-processing involves the preparation of available data (input variables and target classes) for input into MLAs. Data pre-processing methods include the application of corrections such as missing value assignment, mitigating the presence of noise or erroneous measurements, variable extraction and variable selection (Henery 1994b; Guyon 2008).

Data transformations are required to: enhance the relevant signals within input variables; mitigate the effect of inconsistent units; and circumvent the curse-of-dimensionality (Bellman 1961). The curse-of-dimensionality, also known as the Hughes effect (Hughes 1968), relates specifically to the characteristics of Euclidian geometry in high-dimensional spaces, such that the radius of a region varies as the d^{th} root of its volume. This contrast with the distribution of T_a samples within a given region, which varies approximately linearly with the volume (Friedman 1997). Thus, an increase in the dimensionality of variable space with a fixed number of T_a samples leads to these samples being positioned (relatively) further apart. An increase in the sparseness of T_a sample distributions contributes to a decrease in the reliability of estimates of statistical parameters required to construct accurate classifiers (Oommen *et al.* 2008). The curse-of-dimensionality is compounded by inconsistent units between variables. Variable standardisation, e.g. scaling to zero mean and unit variance, offers a simple but robust means of reducing the effects of differences in the relative scale of variables (Wettschereck *et al.* 1997).

Variable extraction is the process by which a specific set of inputs are constructed for a particular application. These may be standard combinations, such as a sum or ratio, which provide a richer source of information than the original variables (Henery 1994b; Kaur & Josan 2011). The act of crafting/extracting informative inputs requires expert knowledge of the intended application as it may not always be obvious which variables are relevant for a given application. The common temptation is to extract as many variables as possible. Nonetheless, incorporating information that is not relevant to a given application may detrimentally affect the ability of the classifier to adequately discriminate between classes (Brazdil & Henery 1994).

Variable selection is the process by which a minimum set of relevant variables are identified from the available input variables and used to train classifiers (Guyon 2008). Variable selection can be viewed as a form of dimensionality reduction as a means of

mitigating the effect of the curse-of-dimensionality. In addition, most if not all MLAs are impacted by the inclusion of redundant (correlated) or irrelevant variables (Wettschereck *et al.* 1997). The selection of a minimum number of relevant variables that best characterise the task at hand can be approached using methods that identify and remove correlated variables using Pearson's correlation coefficients, Principal Component Analysis (PCA) or filtering approaches (Jolliffe 2002; Guyon 2008).

2.2.2.2. Classifier training

As detailed in Section 2.2, in practical situations only the class labels for a limited set of data, T, are known. Therefore, it is necessary to divide T into independent groups for training and evaluation. $\hat{O}_a$ is used to select optimal parameters and train classification models. $\hat{O}_b$ is used to independently evaluate the predictive capabilities of trained classifiers (Witten & Frank 2005; Hastie *et al.* 2009). Otherwise it can be argued that the data used for evaluation has in some way been used to train classifiers, thus violating the assumption of independence (Henery 1994b; Guyon 2009; Hastie *et al.* 2009).

The data in T_a ultimately govern the performance of MLA trained classifiers on new data (Hastie *et al.* 2009). The number of individual T_a samples, the relative number of T_a samples representing individual classes and the number of input variables are factors that influence the statistical distributions of T_a. The statistical distributions of T_a fundamentally influence the resulting decision structures within classification models and hence their performance on unseen data (Henery 1994b). Ultimately, we would like to train a classification model so that it fits $\hat{O}_a$ accurately enough while also being able to generalise well on data not previously seen by the MLA (Burges 1998).

MLA supervised classifiers can be sensitive to different numbers of samples representing individual classes within T_a (Japkowicz & Stephen 2002). Class imbalance increases the risk of classifier overfitting because where there is a large contrast in the number of T_a samples for individual classes over-represented classes become the focus of the trained classifier (Henery 1994b; Japkowicz & Stephen 2002; Wang & Yao 2012). Appendix A documents experiments that assess the sensitivity of different MLAs to imbalanced T_a and T_b class distributions using randomised input variables. The results of these experiments indicate that NB, ANN and SVM are more sensitive to imbalanced class distributions than kNN or RF. The most common approach to addressing classifier sensitivity to imbalance T_a class distributions is to assign high weights to underrepresented classes, thus forcing the

classifier to concentrate on correctly classifying these classes. Nonetheless, prior knowledge of the expected cost of misclassification for all classes in T_a is required when assigning weights (Witten & Frank 2005; Wang & Yao 2012).

If left unconstrained (unregulated) and supplied with adequate information, many MLAs have the ability to fit decision structures to $\hat{O}_a$ without error. However, training MLAs that perfectly fit $\hat{O}_a$ does not generally produce accurate predictions for unseen data, i.e. they are over-fitted. This is because the distribution of $\hat{O}_a$ may not mimic the distribution of all data (Hastie *et al.* 2009). Conversely, underfitting classification models does not adequately characterise the structures within $\hat{O}_a$ due to insufficiently complex decision structures (Henery 1994b; Kuncheva 2004). Therefore, a trade-off exists between inducing the simplest possible classification model that fits $\hat{O}_a$ sufficiently well but which is also general enough to generate accurate predictions for new data (Schölkopf 2003; Guyon 2009).

The selection of appropriate parameters for a given application using available data is an important and non-trivial element of MLA classifier training (MacKay 2003). Most, if not all, MLAs have a range of algorithm specific parameters that require optimisation. The selection of parameters is usually conducted using methods such as cross-validation or bootstrapping. These methods divide $\hat{O}_a$ into subsets containing samples for classifier training (training data) and model parameter adjustment and another subset (validation data) to estimate the performance accuracy (or error) of trained classifiers (Figure 2.6; Henery 1994b; Kuncheva 2004; Guyon 2009; Hastie *et al.* 2009).

Cross-validation is a widely used, simple but objective data-driven method for selecting optimal algorithm parameters for a given application and for assessing the predictive accuracy of trained classification models (Kohavi 1995; Witten & Frank 2005; Hastie *et al.* 2009), especially in circumstances where a limited number of T is available (Henery 1994b; Tan *et al.* 2006; Guyon 2009). During k-fold cross-validation, $\hat{O}_a$ is recursively split into k mutually exclusive and equally sized subsets. For each k fold, one of the k subsets is used as T_b and k 1 subsets are used as T_a. Cross-validation accuracy, which is an estimate of the performance of a trained classification model, is generated by summing or averaging over the results obtained on the k folds (Guyon 2009). 10-fold cross-validation is the recommended method for model parameter selection as it generates stable (low variance) results (Kohavi 1995; Guyon 2008). The algorithm dependent parameters that are selected

are then used to train a final classification model using all samples within $\hat{O}_a$ for use on new data (Hastie *et al.* 2009).

An alternative to *k*-fold cross-validation is the leave-one-out cross-validation where $k = n$ and n is the total number of samples in T_a. For each fold, classifiers are trained on $n - 1$ samples. Leave-one-out cross-validation has the advantage of utilising a large sample for model training, which generates almost unbiased results (Tan *et al.* 2006). However, it is computationally expensive and the results may have high variance due to model overfitting (Guyon 2009). Another alternative to cross-validation is bootstrapping. The bootstrapping method randomly samples, with replacement, from T_a a number of samples equal to the number of samples in T_a. This equates to approximately two-thirds of the data used for training and the remaining third for validation. As for cross-validation, the final results are reported as an average or sum over multiple bootstrap runs (Efron 1983; Breiman 1996).

The most common search strategies in inductive MLA optimisation are exhaustive (discrete parameters) or grid (discretisation of continuous parameters) searches, where all possible combinations of model parameters are trialled in conjunction with *k*-fold cross-validation or similar methods. However, these types of greedy searches can be computationally expensive, especially with fine discretisation of parameter space (Guyon 2009). Nested search methods offer an alternative whereby increasingly smaller nested ranges of parameter values are identified over multiple iterations (Oommen *et al.* 2008).

2.2.2.3. Prediction evaluation

As documented above, $\hat{O}_b$ should be independent to T_a so that a direct and statistically robust indication of the expected performance of a trained classifier is obtained (Guyon 2009). In the case of MLA supervised classifiers, the resulting categorical predictions can be assessed using a confusion (or error) matrix. A confusion matrix represents the classification model performance in tabular form (Congalton & Green 1998). The number of rows and columns in the matrix are equal to the total number of classes in T. The values in each cell of the matrix represent the counts of class predictions for the samples associated with a particular class. Confusion matrix cell entries (Figure 2.7) can be viewed of in terms of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

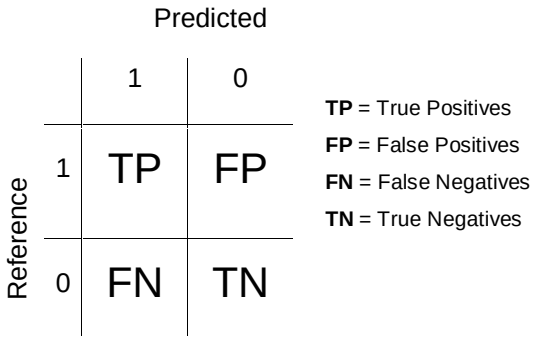

Figure 2.7 Confusion (error) matrix for binary classifications showing the relationships between True and False, Positives and Negatives.

Accuracy and Cohen's kappa statistic (Cohen 1960) are commonly used to evaluate classifier performance (Congalton & Green 1998; Lu & Weng 2007). In these instances, categorical predictions are treated as either correct or incorrect (Friedman 1997). Accuracy, which is simply the fraction of correctly classified $\hat{O}_b$ samples, is defined as:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad [2.14]$$

where TP + TN + FP + FN is the total number of samples and TP + TN represents the correctly classified ($y = \hat{y}$) samples. Accuracy weights the correct prediction of all samples equally regardless of their class (Witten & Frank 2005; Tan *et al.* 2006; Kovacevic *et al.* 2009). This apparent classification accuracy (1 error) is assumed to be an estimate of the probability of a correct classification, from which confidence intervals can be calculated (Kuncheva 2004; Foody 2009).The kappa statistic is calculated using a confusion matrix and represents a measure of the agreement between predicted and observed classes in $\hat{O}_b$ while correcting for agreement that occurs by chance (Witten & Frank 2005). Benchmark performance categories for kappa values suggested by Landis & Koch (1977) include "fair" (0.21–0.40), "moderate" (0.41–0.60), "substantial" (0.61–0.80) and "almost perfect" (0.81–1.00).

For multiclass problems a confusion matrix provides an additional means of identifying in which classes misclassifications, i.e. FP and FN, have occurred (Kuncheva 2004). For each class the fraction of correctly classified $\hat{O}_b$ samples,

$$recall = \frac{TP}{TP+FP}, \quad [2.15]$$

and fraction of correct T_b predictions,

$$precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad [2.16]$$

can be obtained. Recall, also known as Producer's Accuracy, sensitivity and errors of omission, indicates the probability that reference samples for a given class will be correctly classified. Precision, also known as User's Accuracy, positive predictive value and errors of commission, represents the probability that predictions of a given class are correct (Congalton & Green 1998; Witten & Frank 2005).

Most MLAs have the ability to generate, in conjunction with discrete class labels, a vector p_c equal in length to the number of classes in T_a, which represents class membership probabilities. The vector p_c quantifies, for a given sample i, the probability that a class is the correct class given the information in T_a. The maximum class membership probability across all possible classes for a given $\hat{O}_b$ sample constitutes the predicted class. Class membership probabilities provide an opportunity to realistically assess the likelihood of individual predictions and alternative measures for classification model performance assessment such as the *bias*^{2}*-variance* decomposition (Witten & Frank 2005).

The *bias*^{2}*-variance* decomposition (Figure 2.8) separates the prediction error of a learned model, given a fixed range and $\hat{O}_a$ size, into the sum of three non-negative numbers:

$$error = noise + bias^2 + variance. \qquad [2.17]$$

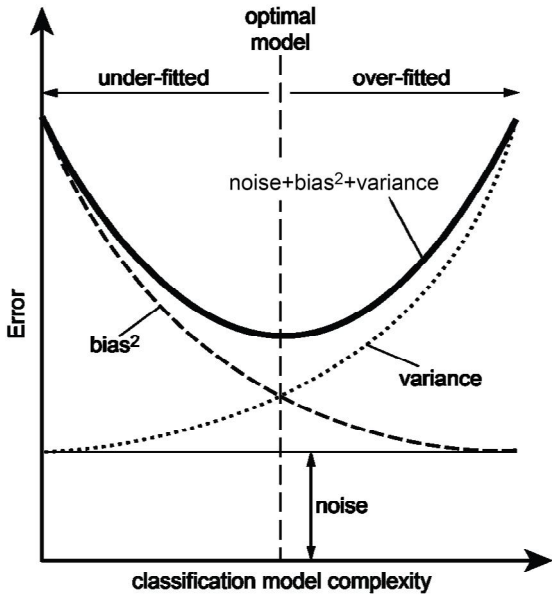

Figure 2.8 Schematic diagram representing the effect of noise and the bias^2*-variance decomposition and its relationship to classifier complexity (under-fitted and over-fitted models) and error.*

This decomposition of error was originally developed for the analysis of numerical outputs by Geman *et al.* (1992). The reducible error of this expression can be divided into the last two terms, $bias^2$ and *variance*, which comprise the error that is a function of the trained classifier (Kohavi & Wolpert 1996; Webb 2000; James 2003). *Noise* represents the irreducible error or the intrinsic noise contained within variables, which is the lower bound on the expected cost of any classification model. *Noise* is defined as (Kohavi & Wolpert 1996):

$$noise = \frac{1}{2}(1 \quad \sum_c a_c{}^2), \qquad [2.18]$$

where a_c represents the underlying probability distribution that, for a given $\hat{O}_b$ sample, the assigned class label is correct. For most applications the class label is assumed to be correct (i.e. 1) and all others are 0.

In practice, *noise* is difficult to estimate as the underlying class probability distributions are not known and there are often too few samples to estimate this reliably (Kohavi & Wolpert 1996; Webb 2000). In comparative studies, *noise* is constant for all learning algorithms and it is therefore independent of the relative performance of MLAs (Kohavi & Wolpert 1996). However, this assumption can lead to an overestimation of $bias^2$ as any variability in the target class is added to this term (James 2003). Nonetheless, as Webb (2000) points out, noise is invariant across the classification models applied to the problem. If the main aim of an experiment is to compare methods, it is not treated as a significant factor in classifier evaluation.

$Bias^2$ is a measure of the degree to which p_c for all possible classes matches that of the target class and is defined as (Kohavi & Wolpert 1996):

$$bias^2 = \frac{1}{2}(\sum_c [a_c \quad p_c]^2). \qquad [2.19]$$

We seek to select an optimal classifier that minimises the $bias^2$ calculated for all $\hat{O}_b$ samples. The optimal outcome is for the classifier to assign the highest probability to the correct class. The closer the maximum class membership probabilities are (for the correct choice) to 1, the smaller the resulting $bias^2$ will be. Therefore, $bias^2$ highlights classifiers that confidently obtain a correct (or incorrect) prediction and is useful for identifying classifiers that are underfitting (Kuncheva 2004; Witten & Frank 2005).

Variance is a measure of the sensitivity of a classification model to $\hat{Q}_a$ and is given as (Kohavi & Wolpert 1996):

$$variance = \frac{1}{2}(1 \quad \sum_c p_c{}^2). \qquad [2.20]$$

Variance increases as the classifier becomes increasingly sensitive to changes in $\hat{Q}_a$ and is useful for identifying overfitting (Kuncheva 2004). Unlike *noise* or $bias^2$, *variance* does not require any knowledge of the true class label for a given sample.

2.3. Unsupervised clustering

Unsupervised clustering embodies machine learning strategies for exploring data with no prior knowledge of inherent divisions/class associations. Unsupervised learning is useful for identifying and visualising natural groups (clusters) within data (Rohwer *et al.* 1994; Ripley 1996; Witten & Frank 2005). Samples within a given cluster should, by definition, be similar to each other, while samples associated to different clusters should not (Kuncheva 2004; Xu & Wunsch 2005; Marsland 2009). In essence, this is equivalent to identifying groups of relatively homogeneous samples (Backer & Jain 1981). The outcomes generated by unsupervised clustering algorithms are generally used to aid exploratory data analysis and the visualisation of associations between data. Hence, successful unsupervised learning generates clusters that provide insight into the structure of data, which can be directly interpreted by domain experts (Ripley 1996). Unsupervised clustering can also be used to identify noisy (corrupted or erroneous) data (Marsland 2009) or to confirm suitable classifications resulting from supervised methods (Ripley 1996).

2.3.1. Clustering strategies

In this section, several well-known and commonly used unsupervised MLA clustering strategies are described: partitioning algorithms, hierarchical algorithms and Self-Organising Maps.

2.3.1.1. Partitioning algorithms

Partitioning clustering algorithms iteratively divide samples into a predetermined number of groups or clusters (Ripley 1996; Witten & Frank 2005). The *k*-means (or *c*-means) clustering algorithm is a simple, effective and widely used method for separating data into *k* groups. For each pre-assigned number of clusters, *k*-means randomly generates a cluster centre and then iteratively adjusts the location of this centre in Euclidean variable space

with respect to the samples that are deemed to be associated with this cluster (Ripley 1996; Witten & Frank 2005; Marsland 2009). This is achieved by minimising the sum of the squared distances from each sample $\mathbf{x}_i$ to its cluster centre m_{ki}:

$$min_{km} \sum_i \|\mathbf{x}_i \quad m_{ki}\|^2. \qquad [2.21]$$

Assigning samples to one of the possible cluster centres using the minimum residual distance principles described above means that there is no guarantee that a global minimum will be obtained (Witten & Frank 2005; Marsland 2009).

The *k*-means algorithm can handle only continuous variables, although the use of medoids can be implemented to cluster categorical data (Ripley 1996). The final clusters are sensitive to the initial, randomly assigned location of cluster centres such that different cluster arrangements will result from slightly different "seed" cluster centres (Witten & Frank 2005). In addition, the *k*-means algorithm is susceptible to data outliers (noise), although this can be circumvented by employing a median rather than mean to identify cluster centres (Marsland 2009).

The *k*-means algorithm described above is a "hard" clustering method such that samples are assigned to one cluster (Figure 2.9). In contrast, the fuzzy *c*-means clustering algorithm (Bezdek 1981) is a "soft" clustering method. Fuzzy *c*-means utilises fuzzy logic allowing for varying degrees of membership for each sample. Cluster membership is quantified via

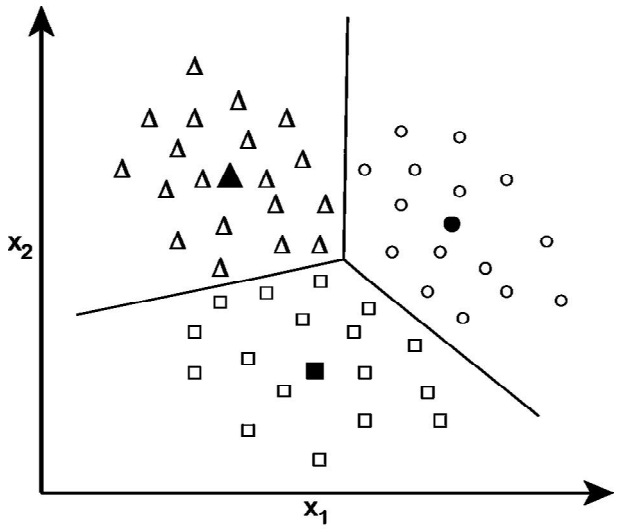

Figure 2.9 Example of k-means (k = 3) clustering in 2D variable space. Filled symbols represent cluster centroids and lines indicate boundaries between clusters.

the assignment of cluster membership probabilities. Cluster membership probability is a vector of length equal to the number of defined clusters that sums to 1. The fuzzy *c*-means approach finds cluster centres, m, from the weighted mean of all samples $\mathbf{x}_i$ by minimising (Ripley 1996):

$$min_w \sum_i \sum_j \boldsymbol{w}^2 \|\mathbf{x}_\mathrm{i} \quad m_v\|^2, \qquad [2.22]$$

where $\boldsymbol{w}$ are the weights of all samples.

2.3.1.2. Hierarchical algorithms

Hierarchical clustering algorithms incrementally generate a 1D dendrogram or tree representing the similarities between clusters. Dendrograms can be constructed either using bottom-to-top (agglomerative) or top-to-bottom (divisive) methods. The agglomerative method starts with *k* clusters that are reassigned at each level of the dendrogram as two existing clusters are iteratively merged. In the divisive approach, the process is initialised with one cluster and is recursively split into two groups until this is no longer possible. Dissimilarities between samples within clusters are specified by the branch levels of a dendrogram (2.10). Groups of clusters are formed by providing users with the option of pruning the dendrogram at some level (Ripley 1996; Witten & Frank 2005).

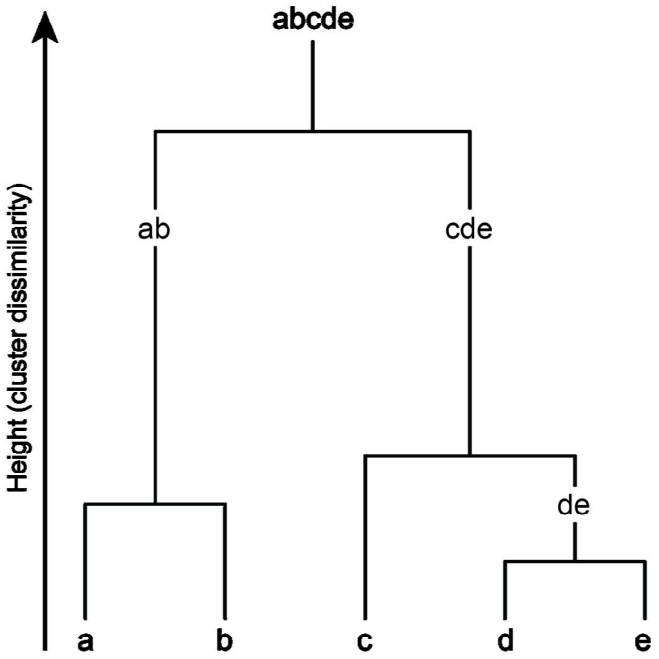

Figure 2.10 Example of a dendrogram with five clusters as output from hierarchical clustering, similar clusters reside on the same branch of the dendrogram. Pruning the dendrogram at a given level will result in similar clusters being merged.

Agglomerative methods for constructing hierarchical clusters simply pick two clusters with the lowest measured dissimilarities and merge them. There are several options for defining the dissimilarity between merged clusters. These are single-linkage, complete- linkage and average-linkage methods. Single-linkage methods join two clusters at a given level if links can be found that joins pairs of samples with dissimilarities less than the specified level. This tends to generate long and loosely connected clusters as only one link is required. Complete-linkage clustering merges two clusters if all samples of one cluster are proximal to the samples of the other cluster. This approach produces compact clusters where relatively similar samples can remain separated up to high levels of the dendrogram. In contrast, the average-linkage approach considers the combined dissimilarity of two clusters to be the average of all dissimilarities between the samples within each cluster, which is dependent on the scale of the dissimilarities. This tends to generate dendrograms with cluster partitions somewhere between those generated by single- linkage and complete-linkage methods (Ripley 1996).

Divisive approaches to hierarchical clustering concentrate on splitting samples at high levels of the dendrogram and are likely to produce more rational groups of data especially with small numbers of clusters. Divisive methods must consider 2^{n-1} 1 partitions on *n* samples into two non-empty groups. As this is computationally infeasible for large *n* only small portions of these partitions are considered. Divisive hierarchical clustering is similar in structure to DT (Ripley 1996).

2.3.1.3. Self-Organising Maps

Self-Organising Maps (SOM), developed by Kohonen (1982) and Kohonen (2001), are akin to an unsupervised form of ANN (Ripley 1996; MacKay 2003). SOM are a constrained version of *k*-means clustering methods in that samples are assigned to the nearest cluster centre. In contrast to *k*-means, SOM attempts to assign a topological structure to groups of similar samples (Hastie *et al.* 2009). SOM achieves this using vector quantisation and measures of vector similarity to map multi-dimensional input data onto a 2D map (2.11). The topological arrangement of cells (nodes) in this 2D map represents the relative proximities of clusters identified within the data (Rohwer *et al.* 1994; Ripley 1996; Kohonen 2001; Marsland 2009).

Training SOM nodes involves an iterative two-stage process. The first stage examines input samples with a predefined number of randomly seeded nodes (seed-nodes). Input samples are compared only to those seed-nodes within a particular radius and assigned to the most similar seed-node based on a measure of vector similarity such as Euclidian distance. The properties of seed-nodes are adjusted to more closely resemble those of the input samples deemed to be represented by the associated seed-node. The second step involves adjusting the properties of proximal seed-nodes to the selected seed-node to again resemble those of the input sample being assessed. These steps are repeated while reducing the radius of assessment and the percentage adjustment of seed-node properties over all input samples (Bierlein *et al.* 2008; Marsland 2009). In this way, SOM spatially constrains proximal seed-nodes such that they become nodes exhibiting the characteristics of clusters of input samples in variable space (Fraser & Dickson 2007; Wehrens & Buydens 2007; Bierlein *et al.* 2008).

In practice, a large number of SOM seed-nodes are chosen and arranged on a 2D hexagonal or rectangular grid. A general rule of thumb is to start with $5 \times \sqrt{n}$ seed-nodes (Vesanto & Alhoniemi 2000). Visualisation of the topological structure of trained SOM nodes using unified-distance matrices (U-Matrix, Ultsch & Vetter 1994) or Principle Component plots (Vesanto 1999) can be used to identify an optimal number of seed-nodes (Vesanto & Alhoniemi 2000). Furthermore, these visualisation methods allow for the formulation of interpretations regarding the significance of SOM node relationships with

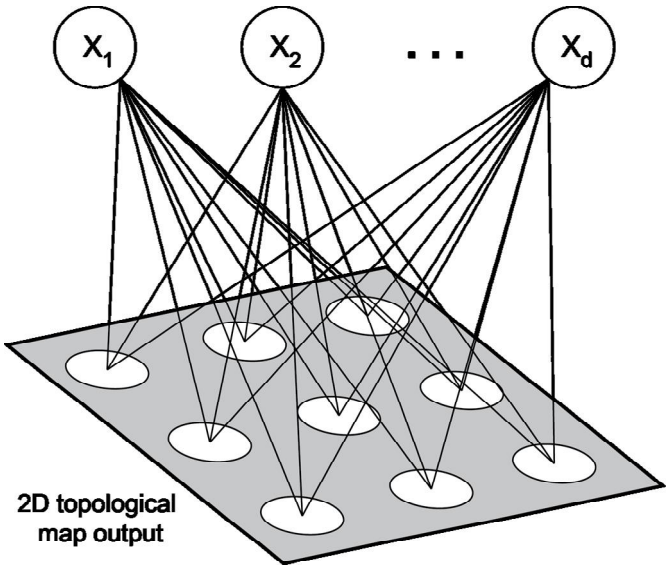

Figure 2.11 Schematic diagram of SOM unsupervised clustering algorithm structure, after Bierlein et al. (2008).

respect to the structures present within the data (Vesanto 1999; Fraser & Dickson 2007; Bierlein *et al.* 2008).

2.3.2. Unsupervised clustering implementation

There are several unsupervised clustering algorithms available. These algorithms all encompass methods that assign *n* samples to *k* clusters. For practical applications, a fundamental question that needs to be addressed when implementing unsupervised clustering algorithms is how many clusters are required to adequately segment data into natural groups (Kuncheva 2004; Xu & Wunsch 2005).

Unlike supervised learning, where the effectiveness of an algorithm to link variables to classes can be estimated using $\hat{O}_b$, there is no systematic approach to evaluating the success of unsupervised clustering outputs (Hastie *et al.* 2009). The main challenge to implementing unsupervised clustering algorithms, which coincidently is the primary influence users have over clustering outputs, is the need to set a predefined number of clusters in which to group available data. In most situations, one does not know the optimal number of clusters to seed. A robust but computationally expensive approach to this problem is to trial different numbers of clusters and evaluate which of these generate the most compact and inclusive groups (Witten & Frank 2005). An alternative approach is to generate a large number of clusters and then merge these based on a criterion that evaluates cluster membership (Gonçalves *et al.* 2008).

2.4. Conclusions

This chapter has detailed the theory of machine learning supervised classification and unsupervised clustering algorithms. There are five general machine learning strategies for supervised classification (Kotsiantis 2007): statistical learning algorithms, e.g. NB and BT; instance-based learners, e.g. kNN; logic-based learners, e.g. DT, RF and BT; Support Vector Machines; and Perceptrons, e.g. ANN. These strategies each have different approaches and assumptions governing their use in practical applications. NB assumes that for a given class the input variables are independent, whereas, kNN, SVM and ANN require input variables to be standardised to a common scale. Despite their differences, all supervised MLAs are fundamentally dependent on the relevance of information within input variables and prior knowledge of the problem, represented by *T*, which is used to train and evaluate classifiers.

The implementation of MLAs for practical supervised classification applications involves three key stages: (1) data pre-processing; (2) classifier training; and (3) prediction evaluation. Data pre-processing provides users with the opportunity to prepare available variables such that they contain information relevant to the intended application. When training MLAs, one or more algorithm specific model parameters require selection in order to optimise classifier performance given the conditions imposed. The selection of parameters is often approached using methods such as *k*-fold cross-validation on a subset of labelled samples, $\hat{O}_a$. An unbiased evaluation of the ability of MLAs to classify new samples is achieved using $\hat{O}_b$. Statistical measures of classifier performance, such as accuracy and the kappa statistic provide an indication of the probability that subsequent classifications will be correct.

Unsupervised clustering machine learning methods are useful for identifying natural groups (clusters) of similar samples. MLA clustering methods offer users the opportunity explore and visualise cluster membership as a means of interpreting patterns within data. In this chapter, I have described three commonly implemented clustering strategies: partitioning algorithms; hierarchical algorithms; and SOM. Fuzzy partitioning methods, i.e. fuzzy *c*-means, allow for varying degrees of cluster membership. Hierarchical methods provide easily interpretable visualisations of cluster similarities via dendrograms. SOM combines the advantages of both partitioning methods and hierarchical methods, such that complex cluster associations can be formed in conjunction with methods for the visualisation and interpretation of cluster topological relationships.

This chapter has deliberately concentrated on MLA theory and implementation with respect to data that are not linked to temporal or spatial frames of reference, as is commonly the case with geoscience data. The temporal and spatial constraints of geoscience data present challenges to the application of MLAs. Chapter 3 builds upon the insights gained in this chapter by placing particular focus on the review of previous research exploring the use of MLAs for the classification of geoscience data.

CHAPTER 3 – A REVIEW OF MACHINE LEARNING FOR GEOSCIENCE CLASSIFICATION APPLICATIONS

Chapter 2 detailed the theoretical background and implementation of supervised and unsupervised machine learning algorithms. This chapter concentrates on reviewing recent peer-reviewed literature concerned with the practical application of machine learning algorithms for real-world data classification problems. Initially, I provide brief a summary of the broad range of non-geoscience data classification applications machine learning algorithms have been used for. This perspective forms a platform for the review of machine learning algorithm use in geoscience data classification applications. Particular focus is placed on the classification of geospatial data, i.e. data that is constrained by a 2D spatial reference frame. I conclude with several key findings and avenues for future research regarding the use of machine learning algorithms for geoscience classification applications.

3.1. Machine learning non-geoscience applications

The primary motivation for the use of machine learning algorithms (MLAs) for data inference is that structured information (patterns) within high-dimensional data can be identified, exploited and subsequently interpreted. This includes the notion that MLAs have the ability to integrate large numbers of seemingly disparate variables in a meaningful and efficient way. This, in turn, provides opportunities for human analysts to formulate interpretations that may not have been obvious using traditional data inference methods.

Naïve Bayes (NB) is a statistical learning algorithm that employs Bayes Theorem to calculate the joint probabilities that a sample is one of *c* classes given a vector inputs (John & Langley 1995; Sivia 1996; Domingos & Pazzani 1997). Despite its naïve assumption that variables are conditionally independent for a given class, NB has been shown to be competitive against many other classification algorithms across a wide range of practical applications. For example, Domingos & Pazzani (1997) evaluated NB against several classification algorithms including the C4.5 Decision Trees (DT) algorithm across 28

datasets obtained from the University of California, Irvine (UCI) data repository (Frank & Asuncion 2010). Dumais *et al.* (1998) compared NB with Bayesian Networks (BN) and linear (kernel) Support Vector Machines (SVM) for a text recognition problem. There was little difference in the accuracy rates of NB and BN, although, SVM performed considerably better than either NB or BN (Dumais *et al.* 1998). The results of a study conducted by Shi & Liu (2011) found that NB was more resilient to increased amounts of missing data (> 10 %) than SVM across the majority of the 24 UCI datasets used in their experiment. In general, NB has been shown to perform well in situations where missing data are present (Michie *et al.* 1994a; Shi & Liu 2011) and where input variables are conditionally independent given the class, such as in medical datasets (Michie *et al.* 1994a; Al-Aidaroos *et al.* 2012).

Despite its simplicity, *k*-Nearest Neighbours (kNN) has been successfully applied to a large number of binary and multiclass classification problems such as: electrochemical waveform analysis (Pichler & Perone 1974); image classification (Brazdil & Henery 1994; Hastie *et al.* 2009); the analysis of marketing and advertising data (Govindarajan & Chandrasekaran 2010); and signal processing (time-series data, Illa *et al.* 2004; Lee *et al.* 2012). These studies indicate that kNN performs as well or better than statistical learning algorithms and DT in situations where the scaling of variables is not important. However, where the dimensionality of data is high, processing time can be considerable for kNN and much longer than some other classifiers (Michie *et al.* 1994a). In addition, the inclusion of irrelevant variables can detrimentally affect kNN performance (Pichler & Perone 1974). More recently, kNN has been advertised as an attractive method for the classification of time-series data. This is because kNN, when applied to time-series data, assesses neighbouring samples (in time) implicitly without the need for complex data transformations. This local approach to learning eliminates the need to apply the often subjective, statistical transformations to time-series data, e.g. sum, mean and variance at discrete time intervals, which in turn mitigates against the loss of information (Lee *et al.* 2012). This trait provides an indication of why kNN performs well in image classification applications (Michie *et al.* 1994a; Hastie *et al.* 2009; Depeursinge *et al.* 2010). The approach to learning implemented by kNN provides it with the ability to implicitly identify samples that, due to the effect of spatial dependency, are spatially proximal without having to carefully construct as series of variables that implicitly represent statistical measures of spatial similarity.

Logic-based algorithms, such as DT and related variants have been used with success in species identification (Cutler *et al.* 2007; De'ath 2007; Morris *et al.* 2007), bioinformatics (Cutler & Stevens 2006; Díaz-Uriarte & De Andres 2006; Yan *et al.* 2012), medical diagnoses (Yang *et al.* 2009; Farhad *et al.* 2012) and image feature identification (Ali *et al.* 2012). In a study conducted by Caruana & Mizil (2006), DT ensemble classifiers, i.e. Boosted Trees (BT) and Random Forests (RF), consistently outperformed other MLAs such as NB, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) on 11 binary class UCI datasets. However, additional processing was required to calibrate BT class membership probabilities in order improve its performance on these datasets (Caruana & Mizil 2006). Banfield (2007) compared DT ensemble classifiers across more than 50 UCI binary class and multiclass datasets. BT and RF were identified as the ensemble DT algorithms that most often obtained statistically significant increases in accuracy over a standard bagging DT (Banfield *et al.* 2007). One of the main reasons that DT (and ensemble DT) algorithms consistently generate highly accurate predictions is they employ a non-parametric approach to learning. This means that DT (and ensemble DT) algorithms do not make assumptions about the statistical distributions of variables and classes (Cutler *et al.* 2007). Moreover, DT (and ensemble DT) algorithms provide users with the potential to learn something about the nature and structure of high-dimensional data with respect to the phenomena under investigation. This is usually achieved by assessing measures of variable importance and plots of partial dependence (Cutler *et al.* 2007; De'ath 2007; Hastie *et al.* 2009).

SVM have been used in applications for financial forecasting and assessment (Shin *et al.* 2005; Trustorff *et al.* 2011); text recognition (Cortes & Vapnik 1995; Joachims 2002; Kaur & Josan 2011); bioinformatics and gene identification (Furey *et al.* 2000; Ding & Dubchak 2001; Hua & Sun 2001; El-Naqa *et al.* 2002); signal processing (Li *et al.* 2003); and medical image classification (Depeursinge *et al.* 2010). These studies have shown that SVM generalises well when faced with small training data, T_a, sizes and high variability/dimensionality within input variables (Trustorff *et al.* 2011). The main advantage of SVM over other algorithms is that they do not define decision boundaries using density estimation; rather they exploit the geometrical characteristics of data by defining decision boundaries using a maximal marginal approach based only on support vectors. In this way, SVM are able to characterise the structures in data using only a

limited number of T_a samples without overfitting (Melgani & Bruzzone 2004; Trustorff *et al.* 2011).

The concept of Artificial Neural Networks (ANN) was first introduced in the 1940s, with the development of McCulloch–Pitts neurons (McCulloch & Pitts 1943). In the 1980s feed-forward back-propagation ANN, known as Multi-Layer Perceptrons (MLP) appeared (Kumar & Thakur 2012). Since then, ANN have been used in a wide range of classification and modelling applications such as: medical diagnosis (Lim *et al.* 1997; Al-Shayea 2011; Åström & Koker 2011; Wu *et al.* 2011; Khan *et al.* 2013); manufacturing technology and process quality control (Huang *et al.* 2007; Mukherjee & Singh 2009; Dêbska & Guzowska-Œwider 2011); ecological modelling (Paruelo & Tomasel 1997; Park *et al.* 2003); and environmental monitoring (Brion *et al.* 2002). Park *et al.* (2003) used a novel approach that combined unsupervised (Self-Organising Maps, SOM) and supervised classification (MLP) ANN algorithms to classifying field sampling sites and predict species richness. In their study, Park *et al.* (2003) employed SOM to initially cluster samples and MLP was then used to classify the resulting clusters into species richness. A large proportion of recent studies investigate more advanced ANN architectures (Kumar & Thakur 2012) such as ensemble schemes (Åström & Koker 2011) and strategies that incorporate fuzzy or probabilistic (Bayesian) estimation of network weights (Wu *et al.* 2011).

3.2. Machine learning geoscience applications

The geosciences encompass a range of scientific disciplines concerned with monitoring and understanding planetary systems, such as geology, geophysics, geochemistry, geodesy, hydrology and metrology. Common to these fields of research is the characteristics of geoscience data, namely the temporal and spatial dimensions from which geoscientific understanding and interpretations are based. The following sections are structured around a review of MLA geoscience classification applications constrained by temporal and spatial reference frames. Particular attention is devoted to the task of supervised and unsupervised lithology classification for the prediction of classes representing spatial distributions of geological materials.

In the previous section, MLA classification applications to general engineering, medical and science based problems were summarised. These studies have, in common, a set of

inputs where the relationships between samples are not bound by a temporal or spatial reference frame. In contrast, geoscience problem are characterised by one or more of these domains, where observations representing samples are positioned in time and/or space. Temporal data are characterised by observations collected at discrete time stamps or intervals. Alternatively, data linked to spatial reference frames contain information on the geographic location of observations. A distinction can be made between geoscience applications based on the number of fundamental physical dimensions explicitly encompassed by available data and the scope of the application. In the following sections, examples of three of these frames of reference are detailed: (1) 0D – data not linked to spatial or temporal dimensions; (2) 1D – one temporal or spatial dimension; and (3) 2D – two spatial dimensions.

3.2.1. Classification of 0D data

Not all geoscience data has to be linked to temporal or spatial reference frames, e.g. geochemical or petrophysical data. The earliest examples of the application of MLAs to 0D geoscience classification problems describe the use of ANN and/or associated variants. Howell *et al.* (1994) employed SOM to cluster and classify asteroids using solar reflectance data spanning wavelengths 0.3–2.5 um with different spectral resolutions. Clusters resulting from the analysis of spectral data with low resolutions were labelled according to end-member asteroid classes interpreted from previous research. When Howell *et al.* (1994) interrogated spectral data with high resolutions using SOM they were able to infer distinct compositional differences between one of the end-member asteroid classes. These results suggested that this particular asteroid end-member class could be subdivided into two distinct categories. In another early study, Sklavounos & Sakellariou (1995) demonstrated the use of MLP for the classification of rock samples from geotechnical data such as: intact rock strength; joint spacing; and orientation; and water content. This research employed a fuzzy classification procedure to represent the degree of membership a given sample had to five rock mass ratings ranging from “Very Good” to “Very Poor” (Sklavounos & Sakellariou 1995).

MLAs have been used for the classification of geochemical data. Lacassie *et al.* (2006) employed SOM to cluster volcanic rock major element geochemical data. The resulting clusters were then assigned class labels representing volcanic rock types, i.e. basalt, andesite, dacite and rhyolite. In a similar study, Das & Iyer (2009) categorised oceanic

basalts using MLP. Savu-Krohn *et al.* (2011) employed SVM to successfully identify the origins of niobium-tantalum ore bodies. SVM have also been used, with reasonable success, to classify soil samples based on physical and chemical properties (Kovacevic *et al.* 2010).

3.2.1. Classification of 1D data

Geological problems that utilise inputs that vary in one physical dimension such as time or depth are bound by 1D reference frames. The bulk of published research regarding the use of MLAs for these types of geoscience applications is dominated by studies into supervised seismic signal (time-series) or drill hole log classification. The basic motivations of the research described in these studies focuses on the use of MLAs for solving complex multivariate classification problems.

3.2.1.1. One temporal dimension

ANN have been used in studies attempting to automate the discrimination of events within time varying signals such as picking the first arrivals of seismic energy (Dai 2003; Hart 2003) and the identification and attenuation noisy or reflections in deep seismic refection surveys (van der Baan & Jutten 2000; Essenreiter *et al.* 2003). These studies concentrated on demonstrating the appropriate use of ANN as tools for approximating continuous functions with arbitrary precision for the detection of discrete events. Van Der Baan & Jutten (2000) describe the process of cross-validation as a means optimising ANN for the selection of parameters controlling network architecture. T_a were identified as the single most important control on the performance of ANN as insufficient examples of the characteristics for each class and overlapping distributions made training accurate classifiers difficult. Van Der Baan & Jutten (2000) further concluded that considerable effort was required to prepare data, train ANN and assess results.

ANN and SVM have both been used to recognise patterns in volcanic tremor data. Masotti *et al.* (2006) used SVM to classify four different eruptive states of Mt Etna from spectrograms of volcanic tremor data. In their study, amplitudes of acoustic waves as a function of time (discrete intervals considered as a continuous time-series dataset) were transformed to Power Spectral Densities and provided as input variables. Labelled samples, *T*, were handpicked and used to train and evaluate a SVM classifier. SVM proved to be highly accurate (> 90 %) based on leave-one-out cross-validation trials (Masotti *et al.* 2006). In a subsequent study, Langer *et al.* (2009) compared SVM and ANN to classify the

same volcanic tremor data as Masotti *et al.* (2006). The results of this experiment indicated that SVM was particularly good at classifying Mt Etna eruptive states. Lower accuracies obtained by ANN were considered to relate to its sensitivity to class imbalanced T_a. Langer *et al.* (2009) acknowledged the need for pre-processing time varying signals into a format, i.e. variance and mean of amplitude and frequency, interpretable by MLAs. In addition, Langer *et al.* (2009) assessed the combined use of supervised classification and unsupervised clustering MLAs for the interrogation of volcanic tremor data. In this instance, SOM was shown to aid the interpretation of sub-classes indicating unique states of volcanic activity at transition zones between the predetermined classes. Langer *et al.* (2009) concluded that combining SVM and SOM obtained an in-depth understanding of the relationships between tremor data and volcanic phenomena.

In a unique study, Ehret (2010) compared the performance of SVM and ANN for the classification of sub-horizontal sub-surface geological boundaries in Ground Penetrating Radar (GPR) two-way-time profiles. *T* were derived from the manual interpretation of two petrophysical properties, electrical conductivity and dielectricity. Input variables contained five different wave-form attributes handcrafted from the GPR traces. Binary SVM and ANN classifiers were employed to identify key contacts between basic lithologies present in the GPR traces, whereas, multiclass classifiers were used to discriminate distinct lithologies. Ehret (2010) found that SVM generated marginally more accurate predictions than ANN for both the binary and multiclass lithology classification problems.

3.2.1.2. One spatial dimension

Drill hole logs of lithology are a crucial element of studies that require information on the sub-surface arrangement of geological materials, e.g. the production of geological models for petroleum and ore deposit exploration (Gifford & Agah 2010; Maiti & Tiwari 2010b). Manual logging of drill hole lithologies is a time consuming and subjective process, often inhibited by poor recovery and limited resources. Therefore, computational tools that provide efficient and robust predictions of classes representing lithologies, such as MLAs, are beneficial in situations where drill hole logs are incomplete (Gelfort 2006; Gifford & Agah 2010; Maiti & Tiwari 2010b). Wireline geophysical data represent measurements of rock physical properties at discrete intervals down a drill hole. These data are obtained either from scans of recovered core or from *in situ* drill hole measurements using wireline instruments. These measurements are assumed to represent petrophysical properties in

immediate vicinity of a drill hole and commonly include magnetic susceptibility, Gamma-Ray Spectrometry (GRS), density and compressional wave velocity (Telford *et al.* 1990).

The most common MLAs to be employed for the classification of wireline geophysical data into lithologies for petroleum industry applications are ANN and associated variants, such as Probabilistic Neural Networks (PNN, e.g., Baldwin *et al.* 1989; Rogers *et al.* 1992; Bhatt & Helle 2002; Koike *et al.* 2002), which use Bayesian methods to estimate network weights (Maiti & Tiwari 2010b; Maiti & Tiwari 2010a). Published research detailing the use of MLAs, other than ANN, for drill hole lithology classification applications commonly involves a comparison with ANN. For example, Wong *et al.* (1995) evaluated Linear and Quadratic Discriminant Analysis (LDA and QDA) and ANN algorithms. ANN were found to be more accurate but required user input to optimise (Wong *et al.* 1995). Schumann (1997) compared LDA, QDA, kNN and ANN for drill hole lithology classification. This study identified kNN as the most accurate algorithm in this situation. More recently, El-Sebakhy *et al.* (2010) concluded that SVM generated substantially more accurate predictions than ANN when classifying seven classes representing carbonate lithofacies. Maiti & Tiwari (2010a; 2010b) assessed the efficacy of PNN to classify three general lithology classes within deep (< 1 km) drill holes. T was obtained by selecting ranges of values representing the physical properties of the three target lithologies from wireline geophysical measurements. Unlike previous research, Maiti & Tiwari (2010b) quantified the uncertainty of PNN predictions and assessed the effect of synthetic correlated noise on classification accuracy. In this case, reasonably accurate predictions were obtained using data with up to 50 % synthetic noise added.

MLAs have been used to discriminate drill hole stratigraphic units within Quaternary sediments. Gelfort (2006) compared a range of algorithms for the classification of drill hole logs including kNN, SVM and PNN. In this dissertation, wireline geophysical data and relative element abundance data, obtained from X-Ray Fluorescence measurements, were used to categorise a relative homogeneous sequence of Quaternary marine sediments into five stratigraphic units. T_a were derived from a random sample of 5 % or 10 % of the total number of samples per class. Gelfort (2006) states that the best classifiers were trained using T_a sampled across the entire interval (drill hole) under investigation. As the aim of this particular study was to compare the performance of several MLAs, T_b represented all other logged data not included in T_a. However, Gelfort (2006) used T_b as a means of optimising SVM classification model parameters, thus, the assumption of

independence between T_a and T_b was violated (Hastie *et al.* 2009). Despite this, Gelfort (2006) identified SVM, PNN and kNN as the most accurate classifiers across two contrasting drill hole datasets. However, the initial classifications contained a large degree of high-frequency noise. In response, Gelfort (2006) employed a majority focal operator (3 m either side of a sample) to smooth predictions, which resulted in a > 10 % increase in accuracy. In an additional experiment, depth (m) was combined with other variables and as the sole variable (Gelfort 2006). Good classification accuracies (~ 80–90 %) were obtained using only depth, whereas, slightly higher accuracies (~ 85–95 %) were generated using all variables including depth. These accuracies were comparable to those resulting from smoothing predictions via majority focal operators to eliminate high-frequency noise. Gelfort (2006) concluded that the number of T_a samples and the spatial distribution of these data are important considerations in terms of classifier accuracy. Furthermore, the inclusion of multiple input variables provided better MLA classifications despite the dominant influence of a small number of key variables (Gelfort 2006).

The studies describe above were primarily concerned with the supervised classification of lithologies from wireline geophysical data. The following studies used both supervised and unsupervised methods to achieve similar outcomes. Toumani (2003) developed a hybrid fuzzy classification system that combined an ANN supervised classifier and an unsupervised clustering algorithm based on the fuzzy *c*-means algorithm (Bezdek 1981). A regularisation parameter, *R*, controlled the degree to which the hybrid classifier coupled supervised and unsupervised learning algorithms. For example, where $R = 0$ (the unsupervised case) the resulting membership assignment was identical to the results of a fuzzy *c*-means classifier. Conversely, where $R = 1$ (the supervised case) the hybrid algorithm will result in class membership probabilities conforming to those predicted by the ANN supervised classifier (Toumani 2003). This hybrid learning algorithm was used to discriminate lithologies within Upper Carboniferous coal bearing sedimentary rocks. These rocks presented a significant challenge to automated classification methods due to the large number of classes (~ 20 in this case) that displayed subtle geophysical differences. Despite considerable mixing of clastic components between lithologies (sandy-siltstone versus siltstone), reasonable separation of three key end-member lithologies (siltstone, sandstone and coal) were obtained (Toumani 2003).

ANN and SOM were investigated by Link & Blundell (2003) as a means of classifying near-surface lithologies from wireline GRS measurements. In this study, ANN was unable

to train a classifier. In response, SOM was used to find six clusters representing the six main lithologies (units with variable amounts of sand, silt and clay) present within the drill hole logs. As in the study conducted by Gelfort (2006), Link & Blundell (2003) employed a majority focal operator to smooth classifications by combining information from the surrounding samples (depth down the drill hole) in order to generate final SOM clusters. Filtered SOM results were plotted and compared to manually interpreted drill hole sections. SOM clusters indicated that the reason why ANN failed to train a classifier was because of the high degree of noise within the GRS data and presence of overlapping statistical distributions between several classes (Link & Blundell 2003). Link & Blundell (2003) concluded that SOM could efficiently provide realistic lithological classifications with minimal user input, thus eliminating the introduction of bias into the lithological classification routine.

Via a completely different approach, Gifford & Agah (2010) combined several MLAs in ensembles of collaborative classifiers to predict geological facies from wireline measurements. This particular study was designed to compare the results of the study conducted by Dubois *et al.* (2007), which evaluated four MLA classifiers, QDA, fuzzy-logic, kNN and ANN using the same input variables and target classes. Dubois *et al.* (2007) had found that ANN generated the most accurate predictions. However, they suggested that a combination of MLAs could exploit the perceived advantages of other non-parametric classifiers. In response to this finding, Gifford & Agah (2010) assessed the performance of combinations of the same or similar (homogeneous) and dissimilar (heterogeneous) algorithms such as NB, kNN, DT, ANN and RF. Diversity between combined algorithms was achieved by supplying different sets of randomly sampled T_a to individual algorithms within an ensemble. The results of this research indicated an increase in overall accuracy, compared to the results obtained by Dubois *et al.* (2007), of up to 6 % was achieved by combining MLAs. In an extension of this research, Gifford & Agah (2012) mapped the presence (or absence) of sub-glacial water from airborne radar transects over the Greenland ice sheet. This study confirmed that combinations of dissimilar algorithms generated accurate classifications. RF was identified as a vital component of the most accurate ensembles (Gifford & Agah 2012).

3.2.1. Classification of 2D data

The classification of data constrained by two fundamental frames of reference is most commonly encountered in image processing. The basic structure of image data is represented by a real-valued matrix covering the scene under investigation. The resolution or scale of these data is restricted by the number of discrete intervals indicated by the frames of reference. Multiple layers of data (variables) can be stored as a stack of matrices, where every cell (pixel) contains a number of values equal to the number of layers in the stack.

Image processing forms the basis of geospatial remote sensing classification applications. The main difference between image data and remote sensing data is that remote sensing images (raster data) are linked to a geographical reference frame. Location on the Earth's surface indicated by the spatial coordinates of pixels. In this section, I review applications designed to classify objects or features located on the Earth's surface (e.g. land cover/vegetation classes). This is followed by a detailed review of use of MLAs for classification problems that attempt to characterise geological materials (i.e. lithological units).

Geospatial classification applications are primarily concerned with the analysis and interpretation of geographically referenced data, such as remote sensing imagery, for the prediction of natural phenomena at or on the Earth's surface. The application of MLAs to raster-based geospatial classification problems is motivated by their ability to handle the typically high-dimensional input data characteristic of remote sensing data. Furthermore, the target phenomena often represent classes that display highly variable spectral signatures with overlapping statistical distributions. Furthermore, it is common that a limited number of observations, T, are available to train and evaluate classifiers (Melgani & Bruzzone 2004; Ham *et al.* 2005; Lu & Weng 2007).

The intended application and spatial extent of available data deemed to be useful for inferring target categories defined the regions of interest for geospatial classification problems. Typical sources of geospatial data include multispectral satellite data such as Landsat TM, Landsat ETM+, Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) imagery and hyperspectral satellite data such as Hyperion and airborne hyperspectral imagery from the HyMap® system. The spatial coverage and spectral and spatial resolutions of these data differ considerably. For example, Landsat TM

data contain seven bands covering the visible and Near Infra-Red (NIR) and Thermal Infra-Red (TIR) wavelengths of the electromagnetic spectrum with a 30 m spatial resolution across the entire globe (Williams 2009). In contrast, HyMap® data is usually available for small regions (100–1000 km^2) and represents more than 100 bands across the visible to Short Wave Infra-Red (SWIR) wavelengths at spatial resolutions of < 5 m, depending on platform altitude (Cocks *et al.* 1998).

3.2.1.3. Land cover/vegetation mapping

The most common geospatial classification applications that MLAs have been employed are for the prediction of land cover or vegetation classes. One of the earliest examples of such an application was conducted by Heermann & Khazenie (1992). In this study, the authors classified multispectral Landsat TM data and assessed the accuracy of ANN classifiers with respect to different numbers of training samples. The absolute number of T_a appeared to be limiting factor with respect to classification accuracy. Furthermore, when faced with a reduced number of input variables (Landsat TM bands) more T_a samples were required to maintain accurate results. Heermann & Khazenie (1992) concluded that ANN was useful for land cover classification from satellite derived multichannel images.

The majority of more recent studies make a distinction between traditional classifiers used in remote sensing, e.g. Maximum Likelihood Classifier (MLC), that assume input variables are normally distributed (parametric) and non-parametric MLAs such as ANN, SVM and DT algorithms. For example, Pal & Mather (2003) evaluated DT against ANN and MLC for the supervised classification of land cover from both multispectral and hyperspectral data. They found that ensemble DT classifiers were best implemented with a small number of input variables, selected using in-built measures of relative variable importance such as the Gini Index (see Section 2.2.1.3). Based on their findings, BT were recommended for problems that contained arbitrarily distributed data. This is because BT, despite generating slightly lower overall accuracies than the other MLAs trialled, were relatively simple to train and required less processing time than the other algorithms. Despite this, MLC is preferred in situations where the data follows Gaussian (or at least unimodal) distributions (Pal & Mather 2003). Nonetheless, Pal and Mather (2003) concluded that the distributions of T_a and test data, T_b, were more important with respect to classifier performance than the choice of algorithm.

More recently, research involving the use of SVM has become more popular. A range of comparative assessments of SVM against commonly used classifiers and other MLAs have been conducted. For example, Pal & Mather (2005) compared SVM to MLC and ANN. The results of this experiment identified SVM as the best performing classifier especially when faced with small numbers of T_a samples and large numbers of input variables. Furthermore, the one-against-one method for combining SVM classifiers for multiclass classification problems performed better that the one-against-all strategy despite an increase in processing time (Pal & Mather 2005). In a similar study, Foody & Mathur (2004) evaluated SVM with respect to ANN, LDA and Classification and Regression Tree (CART) classifiers. These MLAs were employed to classify six vegetation classes from Landsat ETM+ imagery. Non-significant ($p < 0.05$) differences in T_b accuracies were observed between SVM, ANN and LDA. In contrast, the CART classifier consistently performed significantly worse that the other MLAs trialled and required a large number of T_a samples per class to reach equivalent T_b accuracies. Foody & Mathur (2004) concluded that despite SVM sensitivity to different model parameters (ó and C) and the risk of overfitting to T_a when optimising these parameters, it was able to generate the most accurate results when faced with small number of T_a samples, e.g. < 100 per class. Pal (2005) compared RF to SVM for a land cover classification application using Landsat multispectral data. They found that RF performed equally in terms of T_b accuracy to that of the SVM. However, RF was more efficient than SVM and, unlike SVM it offered methods to identify important variables for selection. Furthermore, RF could handle missing data, appeared to be insensitive to imbalanced class distributions and required less user intervention than SVM to optimise parameters (Pal 2005).

The studies summarised above were concerned with the evaluation of MLAs for classifying multispectral remote sensing data. In contrast, Camps-Valls *et al.* (2003) compared SVM, MLP and fuzzy ANN for the classification of soil and crop types from HyMap hyperspectral imagery. They concluded that SVM outperformed the other algorithms trialled when faced with large number of input variables. In addition, SVM classifications did not appear to be detrimentally affected by noisy or irrelevant input variables (Camps-Valls *et al.* 2003). Wilson *et al.* (2004) employed a binary SVM classifier to successfully classify regions affected by heavy metal contamination from HyMap hyperspectral data. Melgani & Bruzzone (2004) reported that a Radial Basis Function (RBF) SVM generated predictions with > 90 % accuracy for a land cover

classification task using hyperspectral data. Their results indicated ~ 10 % improvement in accuracy over ANN, kNN and linear-SVM was obtained using RBF-SVM. Melgani & Bruzzone (2004) suggested that the good performance of SVM was due to the fact that it does not use density estimation to train classifiers, as employed by NB and LDA, Thus, SVM is able to exploit the geometric characteristics of data in variable space by assessing only support vectors (Melgani & Bruzzone 2004). Ham *et al.* (2005) showed, using a vegetation classification example, that the accuracy of RF on spatially disjoint T_b could be improved via methods to select relevant variables from Hyperion hyperspectral imagery. Furthermore, the use of BT was preferred only when large numbers of T_a samples, containing a minimal amount of noise, were available (Ham *et al.* 2005).

Zammit *et al.* (2007) compared kNN and linear, RBF and adapted Spectral Angle Mapper (SAM) SVM kernel classifiers for a binary supervised classification task. This comparison was based on the classification of regions of burnt vegetation from Satellite Pour l'Observation de la Terre (SPOT) multispectral satellite imagery. There was very little difference, i.e. < 1 %, between the T_b accuracies obtained by these classifiers. Nonetheless, Zammit *et al.* (2007) argued that SVM, using the RBF and SAM kernels, generated less False Negatives. Zammit *et al.* (2007) also assessed the use of *k*-means and a SVM classifier coupled with *k*-means unsupervised clustering algorithms. The *k*-means algorithm was used to identify optimal SVM T_a. Despite obtaining non-significant differences in accuracy, compared to the standard RBF SVM classifier, the combined *k*-means and SVM classifier required significantly more processing time to execute (Zammit *et al.* 2007).

Waske & Braun (2009) compared MLC, DT and DT ensemble algorithms (BT and RF) for the classification of nine land cover classes from multi-temporal Single-Aperture Radar (SAR) data. They found that RF significantly outperformed all other MLAs trialled. In addition, the analysis of RF predictions indicated that they were less sensitive to a limited number of T_a and variations in T_a, whilst also being robust to noisy variables (Waske & Braun 2009). In a recent study designed to assess MLA in the context of pixel-based versus object-based classifiers, Duro *et al.* (2012) compared DT, RF and SVM classifiers. Both RF and SVM were found to be more accurate than DT for both the pixel-based and object-based classification trials. Duro *et al.* (2012) concluded there was no statistically significant difference between pixel-based and object-based classifiers using the same MLAs. Pixel-based methods used less input variables and incurred lower overall

processing time to generate outputs, whereas, object-based classifiers generated more visually appealing results, i.e. contiguous regions were classified homogeneously. The major limitation identified with the use of object-based classifiers was that they required considerable user input to formulate appropriate rules in order to generate objects prior to classification (Duro *et al.* 2012).

3.2.1.4. Geological mapping

In this section I concentrate on describing the application of MLAs to the classification of geological materials from geospatial remote sensing imagery. Geological materials include surface/near-surface features that are derived from, or composed of, abiotic materials such as soils, landforms and the parent rocks (lithologies) from which these are derived. This section focuses on the use of MLAs for supervised classification, although, unsupervised clustering and combined (supervised and unsupervised) classification methods are also described.

Supervised classification

The most common MLA used for the supervised classification of lithology for geological mapping applications is ANN. In an early study, An & Chung (1994) used ANN to map four generalised lithologies in a sparsely vegetated high-latitude region from airborne GRS and Total Magnetic Intensity (TMI), Landsat and SPOT data. An & Chung (1994) found that GRS and Landsat data were the most useful variables for discriminating classes representing lithological units with the most difficult to classify samples located close to lithological boundaries (An & Chung 1994). In a similar study, Yang *et al.* (1998) used airborne geophysical, Landsat TM, SAR and SPOT data to classify four distinct lithological classes (carbonate rocks, gneiss metavolcanic rocks, clastic rocks and granodiorite) in a high-latitude region of North America using ANN. GRS data ratios and the SWIR bands of Landsat TM imagery were identified as the most relevant variables for surface lithology mapping applications (Yang *et al.* 1998).

Leverington (2010) applied ANN to a remote sensing lithological classification application in a poorly vegetated high-latitude region of North America. In this study, Leverington (2010) assessed the ability of ANN to discriminate up to 12 classes, including land cover types (snow and green vegetation) and several class representing lithological units such as carbonate rocks, mudstones, sandstones and gabbro. The number of T_a samples for classes was set to 300 except for the gabbro class, which due to its limited spatial distribution was

classified using only 60 T_a samples. In this study, Leverington (2010) compared ANN categorical predictions obtained from Landsat ETM+ multispectral and Hyperion hyperspectral imagery to an interpreted geological map. Although the use of multispectral data generated highly accuracy predictions it was the spectrally distinct classes such as water, snow, green vegetation, mudstones and sandstones that were classified successfully by ANN (Leverington 2010). Furthermore, by simplifying the problem and combining several classes based on common lithologies, i.e. merging sandstones and merging carbonates, Leverington (2010) was able to significantly increase ANN prediction accuracy.

Leverington (2010) attributed misclassifications to the high degree of mixing between classes at the scale of input pixels and spectral similarities between confused classes, either due to concealment of bedrock beneath unconsolidated clastic materials derived from a distal source, or as a result of overlapping spectral properties. Predictions obtained from hyperspectral input variables were displayed as quantitative estimates of the abundance (called spectral un-mixing) of dominant classes and compared to similar outputs using Landsat data (Leverington 2010). For a given pixel, the combined contribution of classes at sub-pixel scales was a function of the heterogeneity of geological materials within a single pixel (Boardman 1989; Leverington 2010). Leverington (2010) concluded that, despite having a low signal-to-noise ratio, the high spatial and spectral resolutions of hyperspectral data were more useful for quantifying class abundances than the low noise and low spatial and spectral resolutions of multispectral data. Nonetheless, the prediction of discrete classes from multispectral imagery provided a sufficient degree of information to identify distinct land cover and geological classes. Given that using Landsat imagery has the potential to generate reasonable results, is widely available and covers the majority of the Earth's surface, Leverington (2010) and Leverington & Moon (2012) suggest that deriving geological maps from Landsat data using standard pixel-based classifiers in sparsely vegetated regions with a high degree of exposed bedrock is a viable option (Leverington 2010).

Grebby *et al.* (2011) compared SOM, implemented as a supervised classifier, to classifications of lithological units generated by MLC. Input variables were taken from high resolution (4 m) airborne multispectral data and a Digital Elevation Model (DEM) derived from Light Detection and Ranging (LiDAR) data. DEM derivatives such as slope, roughness and curvature were generated and integrated. Grebby *et al.* (2011) obtained T_b

accuracies of up to 65 % using SOM and no more than five Principle Component Analysis (PCA) derived input variables. These PCA variables represented 98 % of the variability observed within the original data. As the region under investigation was moderately to heavily vegetated, the introduction of topographic data with spectral data slightly improved classification accuracies (Grebby *et al.* 2011).

The next most common MLA after ANN to be assessed in its ability to predict classes representing lithological units from remote sensing data is SVM. Oommen *et al.* (2008) compared SVM and MLC for the supervised classification of surface lithology from multispectral and hyperspectral satellite imagery. They found that SVM generated predictions with higher accuracies than MLC and was able to effectively utilise a large number of input variables, i.e. 196 hyperspectral bands. In contrast, MLC was not able to construct a classification model for such large numbers of input variables (Oommen *et al.* 2008). Kovacevic *et al.* (2009) applied SVM to the task of lithology classification using multispectral satellite imagery for an arid, poorly vegetated region in Algeria. In this study, classifications were generated from variables representing both geographic (spatial) coordinates and six Landsat ETM+ bands. Waske *et al.* (2009) used Airborne Visible/Infra-Red Imaging Spectrometer (AVIRIS) hyperspectral imagery covering a sparsely vegetated region of Iceland to compare SVM, RF MLC and SAM for the prediction of classes representing Quaternary volcanic units. Waske *et al.* (2009) showed that RF and SVM performed significantly better than MLC and SAM in this situation. Despite RF returning predictions with slightly lower overall accuracies than SVM, it proved to be an attractive method for this type of spatial mapping problem. This is because RF generated robust results with minimal user input despite a limited number of T_a (Waske *et al.* 2009).

Wang *et al.* (2011) compared parallelpiped, Minimum Distance, MLC and so called "intelligent" classification methods of SVM and BT (C4.5 algorithm). In this study, the target classes represented three lithological units in a densely vegetative region of China and input variables were obtained from Landsat data. SVM proved to be the most accurate algorithm of those compared (Wang *et al.* 2011). Wang *et al.* (2011) acknowledged that if alternative remote sensing data to Landsat imagery, such as ASTER or "multi-origin" data (data other than satellite reflectance imagery), are available then more accurate lithology classifications could be obtained where the Earth's surface is concealed by dense vegetation. Yu *et al.* (2012) compared SVM to Minimum Distance and MLC classifiers for

a lithological mapping application in Northern India. They used ASTER and derivatives of the ASTER DEM as input. Variables were combined using a range of contrasting pre-processing methods such as band ratios, PCA and DEM derivatives. They achieved excellent results (> 90 %) for T_b obtained from manually interpreted T, which were confined to outcrop exposures. However, when comparing SVM classifications for the entire scene to a pre-existing geological map a significantly lower accuracy (~ 50 %) was obtained (Yu *et al.* 2012).

Unsupervised clustering

Unsupervised clustering methods provide a means of segregating data into discrete, homogeneous groups of samples. MLAs that offer practical methods for unsupervised clustering usually have some means of allowing users to analyse resulting clusters and thus interpret their geoscientific meaning. For 2D geospatial problems using raster data, clusters should represent contiguous regions with similar characteristics (Fraser & Dickson 2007).

In an early study employing the fuzzy *c*-means unsupervised clustering algorithm, Granath (1988) identified regions of anomalous soil geochemical signatures in the vicinity of the Siljan impact crater. Granath (1988) pre-processed soil geochemical data using PCA in order to exclude anomalous values or outliers from the clustering results. Assessment of the spatial covariance (geostatistical) structure of the fuzzy membership for three clusters was conducted in an attempt to define distinct regions representing anomalous hydrocarbon signatures of deep gas reservoirs. However, the ring like geometry of the impact crater confused the interpretation of isotropic and anisotropic semi-variograms and associated interpolations (Granath 1988). More recently, Wagstaff & Bell (2003) applied the *k*-means clustering algorithm to multispectral data of the Martian surface. The resulting clusters were found to represent interesting surface features such as ice caps but also artefacts of the imaging system. Wagstaff & Bell (2003) concluded that automated analysis methods, i.e. MLAs, offered a means of efficiently handling large volumes of data.

The most widely used unsupervised clustering algorithm for remote sensing geological mapping applications is SOM. Stepinski & Bue (2006) employed SOM to cluster samples obtained from DEM derivatives (slope and roughness), which were then analysed in terms of their context to Martian landforms, i.e. impact craters. In an application that specifically assessed the results of SOM for lithological mapping, Bedini (2009) used HyMap hyperspectral imagery to map seven distinct clusters representing general lithological units.

These clusters were identified to represent lithologies outcropping in a Neoproterozoic carbonatite complex of Greenland. Bedini (2009) indicated that highly accurate agreement (~ 85 %) between SOM clusters and a limited number of *T*, representing field observations, were obtained. The outputs of SOM coupled with the outputs derived from appropriate hyperspectral data pre-processing methods, such as the extraction of mineralogical spectral characteristics, provided additional information for the construction of detailed maps of complex geological terranes (Bedini 2009).

Marsh & Brown (2009) employed SOM to derive clusters, representing seafloor sediments from bathymetry and backscatter data, from a multibeam sonar survey for the production of marine geological maps. In this study, textural derivatives of backscatter data were calculated during pre-processing in order to mitigate the influence of sensor artefacts. More recently, Bedini (2012) mapped hydrothermal alteration zones in the vicinity of a known economic Mo ore deposit using SOM. SOM clusters were assigned to meaningful lithological classes based on what was known about the spatial distribution and spectral characteristics of geological materials in the study area (Bedini 2012). Bedini (2012) also generated maps of the spatial distributions of alteration zones by merging relevant SOM clusters. Carneiro *et al.* (2012) investigated the clustering of airborne geophysics data for geological mapping in the Amazon Basin via SOM. The results of this study identified several distinct groups of clusters representing dominant lithological units. Carneiro *et al.* (2012) concluded that airborne geophysics data provided MLAs with information that could “see” though dense rainforest vegetation.

The studies cited in the preceding paragraphs highlight the need for users to carefully compare unsupervised clustering results with reference data. Reference data commonly includes field observations (Bedini 2009), prior information representing the features of interest (Stepinski & Bue 2006), or comparisons to libraries of hyperspectral reflectance spectra (Bedini 2012). Furthermore, the physical properties and limitations associated with data for remote sensing geological mapping applications must be carefully considered. For example, regions covered by vegetation, interactions with water, regions of shadows cast by high topographic relief and sensor noise introduce artefacts into the results of unsupervised clustering algorithms (Wagstaff & Bell 2003).

Combined supervised and unsupervised methods

The combination of supervised and unsupervised classification methods is most commonly employed where clusters are first identified and then manually labelled by a domain expert. Labelled clusters are then provided as *T* for supervised classification (Langer *et al.* 2009). This approach is closely related to object-based classification methods. However, unlike unsupervised classification, object-based classification methods usually require the provision of user defined rules to initially segment data into homogeneous groups (Duro *et al.* 2012).

Combinations of supervised and unsupervised MLAs have been widely used to map geomorphic features on planetary surfaces. In the first of these studies, Burl *et al.* (1998) describe a machine learning paradigm specifically designed for the identification of volcanoes on Venus. In this study, Magellan radar data, representing the planetary topographic surface was used to extract regions representing homogenous geomorphic elements and *T* were derived from the interpretations of domain experts. Trained supervised classifiers were then used to identify volcanoes in other areas of the Venutian surface (Burl *et al.* 1998). Several other similar studies have used topographic data, such as DEM and associated derivatives obtained from orbital images, to detect and classify manually labelled Martian landforms (e.g., Bue & Stepinski 2007; Stepinski *et al.* 2007; Ghosh *et al.* 2010). Stepinski & Bue (2006) and Stepinski *et al.* (2007) used the *k*-means algorithm to cluster (segment) images of DEM derivatives and spatial coordinates. The inclusion of spatial coordinates ensured that segmented regions were spatially contiguous. The accuracies resulting from classified landform clusters generated by NB, BT (C4.5) and SVM were compared. SVM was found to generate the most accurate and interpretable predictions (Stepinski & Bue 2006; Stepinski *et al.* 2007). In contrast, Ghosh *et al.* (2010) compared watershed segmentation and *k*-means clustering methods as a means of clustering DEM derivatives. In this study, the supervised classification of image segments was conducted using an algorithm that combined NB and BT classifiers. The classifications resulting from this algorithm were compared to those generated by SVM. Detailed landform maps could be constructed by classifying segments obtained from *k*-means clusters with SVM, although *k*-means required considerably more processing time than the watershed segmentation algorithm (Ghosh *et al.* 2010). In a similar study, Dohm *et al.* (2007) used both supervised and unsupervised MLAs to conclude that Martian

mountain ranges were characterised by distinctly different thermal emissions and were, thus, comprised of geological materials of different ages and sources.

Several studies (e.g., Chakraborty *et al.* 2004; Zhou & Chen 2005; De & Chakraborty 2009) have classified backscatter sonar data into generalised seafloor sediment categories. These investigations used either SOM or hybrid vector quantisation classifiers to cluster sonar data into groups representing end-member seafloor sediments. Manually labelled clusters were then used to classify other similar data. These studies showed that pre-processing backscatter sonar data via unsupervised clustering methods lead to significant gains in the accuracy of supervised classifiers. Stumpf & Kerle (2011) investigated the use of combined supervised classification and unsupervised clustering algorithms for a landslide mapping application. This study implemented multi-resolution object-based segmentation methods to cluster samples obtained from a large number of DEM texture derivatives and satellite remote sensing data. RF was then employed to identify the most relevant variables from different textural resolutions and as a binary classifier of image segments (Stumpf & Kerle 2011).

3.3. Practical machine learning implementation

In this section, I summarise the most influential elements of real-world geoscience data and the influence users can have over MLAs to improve their performance. This summary is followed by detailed comments on the specific use of MLAs for practical geoscience applications and potential avenues for further research.

The preceding review of published research into the practical application of MLAs for supervised classification of geospatial data has revealed several salient points that are critical to the performance of MLAs:

1. The quality, dimensionality, spectral resolution and statistical distribution of input variables, especially *T*, are the most important factors with regard to algorithm performance (Heermann & Khazenie 1992; Burl *et al.* 1998; van der Baan & Jutten 2000; Pal & Mather 2003; Foody & Mathur 2004; Lu & Weng 2007; Ehret 2010; Grebby *et al.* 2011; Shaheen *et al.* 2011).

2. Data pre-processing is a time consuming but essential element of MLA implementation. Pre-processing methods are unique for a given application and

the selection of appropriate methods is governed by the nature of target classes and available input variables (Burl *et al.* 1998; Yang *et al.* 1998; Link & Blundell 2003; Ehret 2010; Grebby *et al.* 2011; Wang *et al.* 2011).

3. In situations involving non-linear and non-parametric relationships between multivariate inputs and multiclass targets in conjunction with limited *T*, kNN, ANN, SVM and ensemble DT (RF and BT) generate highly accurate classifications (Pal & Mather 2003; Kuncheva 2004; Lepistö *et al.* 2006; Lu & Weng 2007; Waske & Braun 2009).

4. MLAs require the selection of one or more parameters in order to optimise classifier training for a particular application (Burl *et al.* 1998; Foody & Mathur 2004; Ehret 2010; Kovacevic *et al.* 2010; Yu *et al.* 2012).

5. The use of unsupervised clustering algorithms in conjunction with supervised classification routines provides complementary methods for data pre-processing and the interrogation of relationships between input variables and classification targets (Park *et al.* 2003; Toumani 2003; Zhou & Chen 2005; Dohm *et al.* 2007; Zammit *et al.* 2007; Langer *et al.* 2009; Kraut & Wettergreen 2010).

Of the five points stated above, only one, the quality and characteristics of source data, cannot usually be controlled by the user. This is true except in situations where data is being collected for a specific use. In this instance, the user has some influence over survey design. A pertinent example of the issues associated with data quality is encountered when using geochemical assay data (see Chapter 6). Geochemical data compiled from different surveys using different assay techniques will invariably exhibit inconsistent values between these analyses. Ensuring consistency between individual geochemical samples is achieved by establishing rigorous and systematic assay methods that employ consistent calibration protocols.

The remaining four points, data pre-processing, choice of MLA, parameter optimisation and combination of supervised and unsupervised classifiers, can be explicitly controlled by the user. Apart from the optimisation of MLA parameters, which can be conducted in a semi-automated manner using cross-validation or similar methods (Witten & Frank 2005;

Hastie *et al.* 2009), there does not appear to any strict guidelines regarding the most appropriate MLAs to use for geoscience applications.

The use of MLAs for supervised classification and unsupervised clustering of geoscience data has received increased attention in published literature over the past decade. Typically, geoscience problems involve the use of 1D or 2D, temporally or spatially distributed input variables representing remotely sensed physical properties of the Earth's surface and/or near-surface. Given the temporal and spatial context of geoscience data and inference targets, there are several elements that must be addressed when applying MLAs to these types of problems, namely: the characteristics of data representing geological phenomena; spatial and temporal data pre-processing methods including variable extraction and selection; methods for evaluating predictions and the discovery of interesting relationships between input variables and targets; and the development of integrated workflow routines.

3.3.1. Data

Geoscience phenomena are represented by a wide range of spatial (and temporal) scales. This wide range of scales influences the availability and resolution of these data. In general, higher spectral and spatial resolutions result in improved classification outcomes for the more sophisticated classifiers, e.g. kNN, ANN, SVM and RF. However, a trade-off exists between the coverage, resolution and physical properties that these data represent (Leverington 2010; Grebby *et al.* 2011; Leverington & Moon 2012). For example, the sampling frequencies of wireline geophysical measurements are a limiting factor in the minimum thickness of lithologies that can be discriminated using these data (Toumani 2003).

The use of spectral reflectance data is commonplace in geospatial applications, particularly for land cover classification. The majority of studies that have focussed on the classification of geological materials from multispectral or hyperspectral data have been conducted in regions with little or no vegetation such as high-latitudes, arid zones and extra-terrestrial surfaces (e.g., Yang *et al.* 1998; Dohm *et al.* 2007; Bedini 2009; Kovacevic *et al.* 2009; Waske *et al.* 2009; Leverington 2010; Bedini 2012; Leverington & Moon 2012; Waske *et al.* 2012). This is because these data do not directly measure the properties of geological materials concealed beneath dense vegetation. Therefore, data that "sees" through vegetation, e.g. bare earth DEM, GRS and TMI, are required to improve

lithology classification accuracy in densely vegetated regions (An & Chung 1994; Grebby *et al.* 2011; Wang *et al.* 2011; Carneiro *et al.* 2012; Yu *et al.* 2012).

3.3.2. Data pre-processing

The conversion of geophysical measurements into geological classes is a difficult proposition for standard inference techniques. This is because of the complex non-linear relationships that commonly exist between geophysical observations and rock properties (Ehret 2010). For example, in studies assessing the use of MLAs for 1D temporal data problems, such as the classification of volcanic tremor data (e.g., Masotti *et al.* 2006; Langer *et al.* 2009), raw time-series data was transformed to variables representing frequency components of discrete time intervals. In this way, an implicit notion of temporal context is provided to MLAs.

In 1D and 2D spatial geoscience problems, the use of focal operators (moving windows, convolution filters) to extract derivatives or textural information has been used to provide implicit notions of the variability of physical properties in space (e.g., Grebby *et al.* 2011; Wang *et al.* 2011; Yu *et al.* 2012). Nonetheless, handcrafting variables that provide an implicit notion of temporal or spatial context complicates the inference process. Firstly, considerable effort is required to identify appropriate variables for transformation and the scale with which to assess spatial context. These choices are often arbitrary and based on the subjective knowledge of an expert (Burl *et al.* 1998). It is for this reason that rigorous investigations into appropriate transformation methods and scales of assessment for a given application are required (Lu & Weng 2007). Secondly, the incorporation of additional layers of information will invariably result in the inclusion of irrelevant information, which may detrimentally affect the performance of MLAs. Therefore, variable selection is required to reduce this information to a minimum number of relevant variables (Lu & Weng 2007). Some MLAs, such as RF, are seen as attractive algorithms for classifying high-dimensional input variables as they offer in-built methods for the selection of important variables and/or mechanisms that control how the algorithm learns to focus on relevant variables during classifier training (Wang *et al.* 2011).

3.3.3. Prediction evaluation

Several studies have evaluated the effect of the number of T_a samples on MLA classification accuracy (e.g., Heermann & Khazenie 1992; An & Chung 1994; Ham *et al.* 2005; Pal & Mather 2005; Waske *et al.* 2009; Waske & Braun 2009). In general,

increasing the number of T_a samples lead to improvements in accuracy. However, T_a for 2D geospatial classification applications are usually limited in number and confined to spatially discrete locations. The evaluation of MLA classifiers using T_b that are spatial adjacent to T_a may violate the assumption of independence. In these situations, classifier evaluation may provide an overinflated indication of the accuracy of classifications outside the spatial domain of T_a (Burl *et al.* 1998). An example of this effect was observed in the study conducted by Yu *et al.* (2012). In this study, T representing lithological classes within the study area were derived from handpicked examples observed in Quickbird imagery. The "independent" T_b were sampled randomly from these labelled data. The reported T_b accuracies were significantly higher than the accuracies obtained for classifications that were compared to a pre-existing geological map. In this case, the trained SVM classifier was over-fitted to T_a, which were spatially confined to discrete regions representing outcrop exposures.

In a recent study, Loosvelt *et al.* (2012) assessed methods for estimating pixel-based prediction uncertainty for a vegetation classification task using RF. They concluded that the classification uncertainty was related to interclass similarities. None of the studies reviewed in the previous sections have explored the use of class membership probabilities to assess the uncertainty of predictions. Class membership probabilities may provide the opportunity to characterise interclass similarities commonly encountered between lithological units (Toumani 2003; Gelfort 2006; Leverington 2010).

The interrogation of decision structures induced by MLA has not been thoroughly assessed or examined for geoscience applications. A significant hurdle with respect to assessing the relationships between variables and classes is that many MLAs, such as ANN and RF, are "black boxes". Therefore, it is difficult to comprehend the internal structures induced by these MLAs (Breiman 2001; Hastie *et al.* 2009). Nonetheless, De'ath (2007) and Cutler *et al.* (2007) describe methods that assess the interactions between inputs and targets for ecological modelling applications using ensemble DT algorithms in conjunction with variable importance rankings and partial-dependence plots.

3.3.4. Integrated workflow

It is important to have integrated software infrastructure for data labelling, database support, experimental design and reporting, optimisation, implementation and for the evaluation and visualisation of MLA classifications (Burl *et al.* 1998). Successful MLA

classifications are primarily dependent on the use of high-quality data, robust classification procedures and operator skill and experience. The integration of Geographic Information System (GIS) software with MLA classification workflow is a crucial component of generating robust and interpretable results (Lu & Weng 2007). The benefits of an integrated system for the objective inference of real-world data will be fully realised in situations where massive datasets are being used for data inference (Burl *et al.* 1998; Sellars *et al.* 2013).

3.4. Conclusions

MLAs have been used extensively for supervised classification and unsupervised clustering applications in the fields of finance, medical diagnosis, science and engineering. For 2D geospatial applications, the main focus of the majority of published research investigates the use of MLAs for land cover classification from satellite or airborne multispectral and hyperspectral data. Automated geological mapping from remote sensing data using MLAs has the potential to compliment tasks such as: environmental mapping and monitoring; economic resources assessment and exploration; and the study of surface process and geological histories (Dohm *et al.* 2007; Ding *et al.* 2008; Leverington 2010). To date, there have been a relatively small number of studies that compare MLAs for the supervised classification of lithologies from geophysical remote sensing data. The bulk of research has either focused on comparisons between different data sources (e.g., An & Chung 1994; Kovacevic *et al.* 2009; Leverington 2010), or the comparison of MLAs with more traditional remote sensing supervised classification methods such as MLC (e.g., Leverington & Moon 2012; Yu *et al.* 2012). Some researchers have assessed commonly used MLAs against each other for lithology mapping (e.g., Waske *et al.* 2009; Wang *et al.* 2011) and found that SVM and RF generated highly accurate predictions.

An extensive review of literature concerning the application of MLAs to 2D geoscience classification problems has revealed several avenues for further research. Firstly, data must be applicable to the inference target, e.g. using hyperspectral remote sensing data in regions covered by dense vegetation does not provide direct information on the properties of geological materials. Therefore, investigations are required into the use of airborne geophysical data, which directly observe specific properties of surface/near-surface geological materials. Secondly, the extraction and selection of variables that provide some notion of explicit or implicit, temporal or spatial context to MLAs requires additional

research to formally identify the best approaches. Thirdly, the assumption that T_a is independent of T_b is often violated when assessing for geospatial classifications. Therefore, an assessment of how the spatial distribution of T_a affects the performance of MLAs with respect to the spatial distribution of T_b is required. Fourthly, investigations are needed into methods that assess the uncertainty of MLA predictions and the interactions between input variables and targets. Finally, integrated workflows for data pre-processing, MLA training (optimisation and implementation) and the evaluation of MLA classifications will facilitate efficient and robust inference outcomes of complex geological phenomena.

CHAPTER 4 – GEOLOGICAL MAPPING USING REMOTE SENSING DATA: A COMPARISON OF FIVE MACHINE LEARNING ALGORITHMS, THEIR RESPONSE TO VARIATIONS IN THE SPATIAL DISTRIBUTION OF TRAINING DATA AND THE USE OF EXPLICIT SPATIAL INFORMATION

Published in Computers & Geosciences, vol. 63, pp. 22-33, 2014

4.0. Abstract

Machine learning algorithms are a powerful group of data-driven inference tools that offer an automated means of recognising patterns in high-dimensional data. Hence, there is much scope for the application of machine learning algorithms to the rapidly increasing volumes of remotely sensed geophysical data for geological mapping problems. We carry out a rigorous comparison of five machine learning algorithms: Naïve Bayes; *k*-Nearest Neighbours; Random Forests; Support Vector Machines; and Artificial Neural Networks, in the context of a supervised lithology classification task using widely available and spatially constrained remotely sensed geophysical data. We make a further comparison of machine learning algorithms based on their sensitivity to variations in the degree of spatial clustering of training data and their response to the inclusion of explicit spatial information (spatial coordinates). Our work identifies Random Forests as a good first choice algorithm for the supervised classification of lithology using remotely sensed geophysical data. Random Forests is straightforward to train, computationally efficient, highly stable with respect to variations in classification model parameter values and as accurate as, or substantially more accurate than the other machine learning algorithms trialled. The results of our study indicate that as training data becomes increasingly dispersed across the region under investigation, machine learning algorithm predictive accuracy improves dramatically. The use of explicit spatial information generates accurate lithology predictions but should be used in conjunction with geophysical data in order to generate geologically plausible predictions. Machine learning algorithms, such as Random Forests,

are valuable tools for generating reliable first-pass predictions for practical geological mapping applications that combine widely available geophysical data.

Key words: Geological mapping; remote sensing; machine learning; supervised classification; spatial clustering; spatial information.

4.1. Introduction

Machine learning algorithms (MLAs) use an automatic inductive approach to recognise patterns in data. Once learned, pattern relationships are applied to other similar data in order to generate predictions for data-driven classification and regression problems. MLAs have been shown to perform well in situations involving the prediction of categories from spatially dispersed training data (T_a) and are especially useful where the process under investigation is complex and/or represented by a high-dimensional input space (Kanevski *et al.* 2009). In this study we compare MLAs, applied to the task of supervised lithology classification, i.e. geological mapping, using airborne geophysics and multispectral satellite data. The algorithms that we evaluate represent the five general learning strategies employed by MLAs: Naïve Bayes (NB) – statistical learning algorithms; *k*-Nearest Neighbours (kNN) – instance-based learners; Random Forests (RF) – logic-based learners; Support Vector Machines (SVM); and Artificial Neural Networks (ANN) – Perceptrons (Kotsiantis 2007).

The basic premise of supervised classification is that it requires T_a, containing labelled samples representing what is known about the inference target (Ripley 1996; Witten & Frank 2005; Kotsiantis 2007). MLA architecture and the statistical distributions of observed data guides the training of classification models, which is usually carried out by minimising a loss (error) function (Kuncheva 2004; Marsland 2009). Trained classification models are then applied to similar input variables to predict classes present within T_a (Witten & Frank 2005; Hastie *et al.* 2009).

The majority of published research focusing on the use of MLAs for the supervised classification of remote sensing data has been for the prediction of land cover or vegetation classes (e.g., Huang *et al.* 2002; Foody & Mathur 2004; Ham *et al.* 2005; Pal 2005; Waske & Braun 2009; Song *et al.* 2012). These studies use multispectral or hyperspectral imagery as inputs and T_a is sourced from manually interpreted classes. MLAs such as RF, SVM and ANN are commonly compared in terms of their predictive accuracies to more traditional

methods of classifying remote sensing data such as the Maximum Likelihood Classifier (MLC). In general, RF and SVM outperform ANN and MLC, especially when faced with a limited number of T_a samples and a large number of inputs and/or classes. Previous investigations into the use of MLAs for supervised classification of lithology (e.g., Oommen *et al.* 2008; Waske *et al.* 2009; Leverington 2010; Leverington & Moon 2012; Yu *et al.* 2012) focus on comparing MLAs, such as RF and/or SVM, with more traditional classifiers.

Common to all remote sensing image classification studies is the use of geographical referenced input data containing co-located pixels specified by coordinates linked to a spatial reference frame. Despite this, inputs used in the majority of studies cited do not include reference to the spatial domain. This is equivalent to carrying out the classification task in geographic space where samples are only compared numerically (Gahegan 2000). To date few investigations have evaluated the performance of MLAs in conjunction with the input of spatial coordinates. Kovacevic *et al.* (2009), for example, investigated the performance of SVM using Landsat ETM+ multispectral bands and spatial coordinates, concluding that, given T_a of suitable quality, there was sufficient information in the spatial coordinates alone to make reliable predictions. However, when applying trained classification model to regions outside the spatial domain of T_a the information in Landsat ETM+ bands became increasingly important.

Our supervised lithology classification example evaluates MLA performance in the economically important Broken Hill area of western New South Wales, a region of Palaeoproterozoic metasedimentary, metavolcanic and intrusive rocks with a complex deformation history. In this study, we use airborne geophysics and Landsat ETM+ imagery to classify lithology. Airborne geophysics, unlike satellite spectral reflectance imaging, it is not affected by cloud and/or vegetation cover and represents the characteristics of surface and near-surface geological materials (Carneiro *et al.* 2012). Landsat ETM+ data are freely available and have extensive coverage at medium resolutions over large regions of the globe (Leverington & Moon 2012). Although hyperspectral data has been shown to generate excellent results in sparsely vegetated regions due to high spectral and spatial resolutions (Waske *et al.* 2009) this data is limited in its coverage and ability to penetrate dense vegetation for the characterisation of geological materials (Leverington & Moon 2012).

We explore and compare the response of MLAs to variations in the spatial distributions and spatial information content of T_a and their ability to predict lithologies in spatially disjoint regions. We facilitate this comparison by conducting three separate experiments: (1) assessing the sensitivity of MLA performance using different T_a on test data (T_b) not located within training regions; (2) random sampling of multiple T_a with contrasting spatial distributions; and (3) using three different combinations of input variables, X and Y spatial coordinates (XY Only), geophysical data (airborne geophysics and Landsat ETM+ imagery, No XY) and combining geophysical data and spatial coordinates (All Data). These experiments are combined to provide a robust understanding of the capabilities of MLAs when faced with T_a collected by geologists in challenging field sampling situations using widely available remotely sensed input data.

4.1.1. Machine learning for supervised classification

Classification can be defined as mapping from one domain (input data) to another (target classes) via a discrimination function $y = f(\mathbf{x})$. Inputs are represented as d vectors of the form $\langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d \rangle$ and y is a finite set of c class labels $\{y_1, y_2, \ldots, y_c\}$. Given instances of $\mathbf{x}$ and y, supervised machine learning attempts to induce or train a classification model f', which is an approximation of the discrimination function $\hat{y} = f(\boldsymbol{x})$ and links input data to target classes (Gahegan 2000; Hastie *et al.* 2009; Kovacevic *et al.* 2009). In practice, as we only have class labels for a limited set of data, $\hat{O}$, it is necessary to divide available data into separate groups for training and evaluating MLAs. $\hat{O}_a$ are used to optimise and train classification models via methods such as cross-validation. $\hat{O}_b$ contains an independent set of samples not previously seen by the classifier and is used to provide an unbiased estimate of classifier performance (Witten & Frank 2005; Hastie *et al.* 2009).

The task of MLA supervised classification can be divided into three general stages (1) data pre-processing, (2) classification model training and (3) prediction evaluation. Data pre-processing aims to compile, correct, transform or subset available data into a representative set of inputs. Pre-processing is motivated by the need to prepare data so that it contains information relevant to the intended application (Guyon 2008; Hastie *et al.* 2009).

MLAs require the selection of one or more algorithm specific parameters that are adjusted to optimise their performance given the available data and intended application (Guyon 2009). With only $\hat{O}_a$ available for training and estimating performance, the use of techniques such as *k*-fold cross-validation is required. Trained classification model

performance is usually estimated by summing or averaging over the results of *k* folds. Parameters that generate the best performing classifier, given the conditions imposed, are used to train a MLA using all of the samples in $\hat{O}_a$ (Guyon 2009; Hastie *et al.* 2009).

An unbiased evaluation of the ability of MLAs to classify samples not used during training, i.e. to generalise, is achieved using $\hat{O}_b$ (Witten & Frank 2005; Hastie *et al.* 2009). Classifier performance metrics such as overall accuracy and kappa (Lu & Weng 2007) are easily interpretable and commonly used measures of MLA performance for remote sensing applications (Congalton & Green 1998).

4.1.2. Machine learning algorithm theory

4.1.2.1. Naïve Bayes

NB is a well-known statistical learning algorithm recommended as a base level classifier for comparison with other algorithms (Henery 1994a; Guyon 2009). NB estimates class conditional probabilities by "naïvely" assuming that for a given class the inputs are independent of each other. This assumption yields a discrimination function indicated by the products of the joint probabilities that the classes are true given the inputs. NB reduces the problem of discriminating classes to finding class conditional marginal densities, which represent the probability that a given sample is one of the possible target classes (Molina *et al.* 1994). NB performs well against other alternatives unless the data contains correlated inputs (Witten & Frank 2005; Hastie *et al.* 2009).

4.1.2.2. *k*-Nearest Neighbours

The *k*-Nearest Neighbours (kNN) algorithm (Fix & Hodges 1951; Cover & Hart 1967) is an instance-based learner that does not train a classification model until provided with samples to classify (Kotsiantis 2007). During classification, individual T_b samples are compared locally to *k* neighbouring T_a samples in variable space. Neighbours are commonly identified using a Euclidian distance metric. Predictions are based on a majority vote cast by neighbouring samples (Henery 1994a; Witten & Frank 2005; Kotsiantis 2007). As high *k* can lead to overfitting and model instability, appropriate values must be selected for a given application (Hastie *et al.* 2009).

4.1.2.3. Random Forests

RF, developed by Breiman (2001), is an ensemble classification scheme that utilises a majority vote to predict classes based on the partition of data from multiple decision trees.

RF grows multiple trees by randomly subsetting a predefined number of variables to split at each node of the decision trees and by bagging. Bagging generates T_a for each tree by sampling with replacement a number of samples equal to the number of samples in the source dataset (Breiman 1996). RF implements the Gini Index to determine a "best-split" threshold of input values for given classes. The Gini Index returns a measure of class heterogeneity within child nodes as compared to the parent node (Breiman *et al.* 1984; Waske *et al.* 2009). RF requires the selection of *mtry* which sets the number of possible variables that can be randomly selected for splitting at each node of the trees in the forest.

4.1.2.4. Support Vector Machines

SVM, formally described by Vapnik (1998), has the ability to define non-linear decision boundaries in high-dimensional variable space by solving a quadratic optimisation problem (Karatzoglou *et al.* 2006; Hsu *et al.* 2010). Basic SVM theory states that for a non-linearly separable dataset containing points from two classes there are an infinite number of hyperplanes that divide classes. The selection of a hyperplane that optimally separates two classes, i.e. the decision boundary, is carried out using only a subset of T_a known as support vectors. The maximal margin *M* (distance) between the support vectors is taken to represent the optimal decision boundary. In non-separable linear cases, SVM finds *M* while incorporating a cost parameter *C*, which defines a penalty for misclassifying support vectors. High values of *C* generate complex decision boundaries in order to misclassify as few support vectors as possible (Karatzoglou *et al.* 2006). For problems where classes are not linearly separable, SVM uses an implicit transformation of input variables using a kernel function. Kernel functions allow SVM to separate non-linearly separable support vectors using a linear hyperplane (Yu *et al.* 2012). Selection of an appropriate kernel function and kernel width, ó, is required to optimise performance for most applications (Hsu *et al.* 2010). SVM can be extended to multiclass problems by constructing $\frac{c\,(c-1)}{2}$ binary classification models, the so called one-against-one method, in order to generate predictions based on a majority vote (Hsu & Lin 2002; Melgani & Bruzzone 2004).

4.1.2.5. Artificial Neural Networks

ANN have been widely used in science and engineering problems. They attempt to model the ability of biological nervous systems to recognise patterns and objects. ANN basic architecture consists of networks of primitive functions capable of receiving multiple weighted inputs that are evaluated in terms of their success at discriminating the classes in

$\hat{Q}_a$. Different types of primitive functions and network configurations result in varying models (Rojas 1996; Hastie *et al.* 2009). During training network connection weights are adjusted if the separation of inputs and predefined classes incurs an error. Convergence proceeds until the reduction in error between iterations reaches a decay threshold (Rojas 1996; Kotsiantis 2007). We use feed-forward networks with a single hidden layer of nodes, a so called Multi-Layer Perceptron (MLP, Venables & Ripley 2002) and select one of two possible parameters: *size*, the number nodes in the hidden layer.

4.1.3. Geology and tectonic setting

This study covers an area of ~ 160 km^2 located near Broken Hill, far western New South Wales, Australia (Figure 4.1). The geology of the Broken Hill Domain (BHD, Webster 2004) features an inlier of the Palaeoproterozoic Willyama Supergroup (WSG, Willis *et al.* 1983). WSG contains a suite of metamorphosed sedimentary, volcanic and intrusive rocks, deposited between 1710–1704 ± 3 Ma (Page *et al.* 2005a; Page *et al.* 2005b). WSG features complex lithology distributions resulting from a long history of folding, shearing, faulting and metamorphism (Stevens 1986; Webster 2004). BHD is of significant

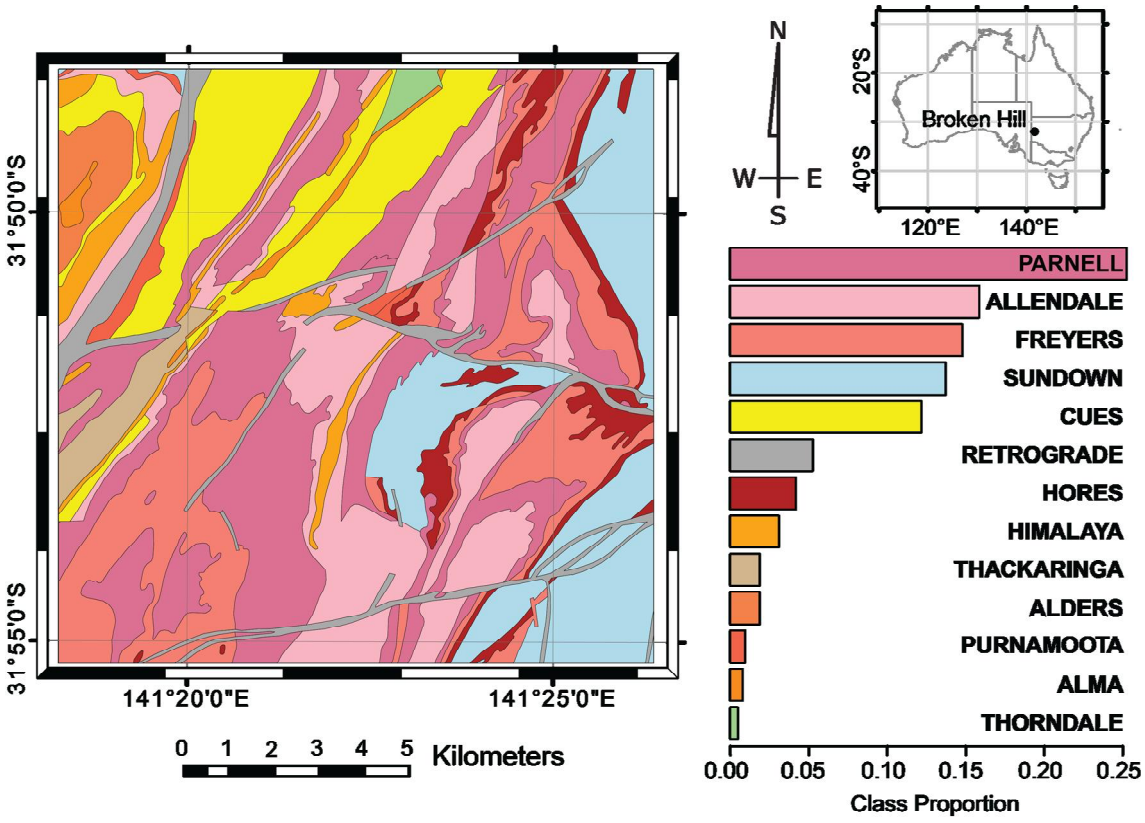

Figure 4.1 Reference geological map, after Buckley et al. (2002) and associated class proportions for the 13 lithological classes present within the Broken Hill study area, modified from Cracknell & Reading (2013).

Table 4.1 Summary of 13 lithological classes within the Broken Hill study region, complied using information from Willis et al. (1983) and Buckley et al. (2002).

Class Label	Class (undiff.)	Class	Description
RETROGRADE		RETROGRADE SCHIST	Retrograde micaceous schist, original stratigraphic position unknown.
SUNDOWN		SUNDOWN GROUP	Predominantly non-graphitic meta-sediment. Pelite-psammopelite units most abundant and more common in the lower half. Psammite-psammopelite units more common in the upper half
HORES	**PURNAMOOTA**	HORES GNEISS	Mainly garnet-bearing quartzo-feldspathic gneiss. Medium to fine-grained quartz-plagioclase-K-feldspar-biotite garnet gneiss
FREYERS		FREYERS METASEDIMENTS	Mainly metasediments ranging from well bedded pelitic/psammopelitic schists, with some psammitic intervals, to psammopelitic or psammitic/pelitic metasediments
PARNELL		PARNELL FORMATION	Extensive bodies of basic gneiss, lenticular masses of garnet-bearing quartzo-feldspathic gneiss and "lode horizon" rocks, intercalated with pelitic to psammopelitic and psammitic metasediments
ALLENDALE		ALLENDALE METASEDIMENTS	Mainly metasediment and metasedimentary composite gneiss. Variable ratio of pelite, psammopelite, psammite. Commonly garnet rich. Sporadic bodies of basic gneiss and quartz-gahnite,
HIMALAYA	**THACKARINGA**	HIMALAYA FORMATION	Extensive bodies of medium-grained saccharoidal leucocratic sodic plagioclase quartz + K-feldspar biotite rocks, with variably developed interbedded metasedimentary composite gneiss and basic gneiss.
CUES		CUES FORMATION	Mainly psammopelitic to psammitic composite gneisses or metasediments, with intercalated bodies of basic gneiss
ALDERS		ALDERS TANK FORMATION	Consists largely of composite gneisses, with little or no basic gneiss, local minor plagioclase-quartz rocks and minor granular quartz-iron oxide/iron sulfide "lode" rocks
ALMA		ALMA GNEISS	Mainly medium to coarse-grained quartz-feldspar-biotite + garnet gneiss with abundant to sporadic megacrysts of K-feldspar and plagioclase, or of aggregates of quartz + feldspar
THORNDALE		THORNDALE COMPOSITE GNEISS	Mainly sedimentary quartz-feldspar-biotite-sillimanite ± garnet ± cordierite composite gneiss, consisting of interlayer psammite and psammopelite, generally with minor pelite and abundant pegmatitic segregations commonly disrupt bedding

economic importance as it is the location of the largest and richest known Broken Hill Type stratiform and stratabound Pb–Zn–Ag deposit in the world (Willis *et al.* 1983; Webster 2004).

Within the study area defined for this experiment are 13 lithology classes that form a chronological sequence younging from west to east. In general, WSG basal units are dominated by quartzo-feldspathic composite gneisses (Thorndale Composite Gneiss and Thackaringa Group) and are overlain by dominantly psammitic and pelitic metasedimentary rocks (Allendale Metasediments, Purnamoota Subgroup and Sundown Group). Table 4.1 provides a detailed summary of lithology classes present within the study area.

BHD deformation history can be summarised into four events (Stevens 1986; Webster 2004). The first two events of the Olarian Orogeny (1600–1590 Ma, Page *et al.* 2005a; Page *et al.* 2005b) are associated with amphibolite–granulite facies metamorphism and

north-northeast to south-southwest trending regional fabric. A third event is characterised by localised planar or curviplanar zones of Retrograde Schist. These zones fulfil the role of faults in the BHD and display well developed and intense schistosity, strongly deformed metasediment bedding and generally displace the units they intersect (Stevens 1986). The fourth deformation event, associated with the Delamerian Orogeny (458–520 Ma, Webster 2004), is interpreted from gentle dome and basin structures.

4.2. Data

The published 1:250,000 Broken Hill digital geological map, compiled by Buckley *et al.* (2002), was used to obtain labelled samples representing lithology classes for training and to evaluate MLA predictions. We maintained the number of T_a at 10 % (~ 6500) of the total number of samples in the selected scene. Multiple sets of T_a were randomly sampled from approximately circular regions, randomly seeded across the study area. The number of T_a regions were varied from 1–1024, such that the number of regions was equal to 2^a, where a represents sequential integers from 0–10. In all cases, the total coverage of training regions equates to > 10 % and < 20 % of the study area (Figure 4.2).

Spatially disjoint T_b samples were sampled from all other pixels in the study area not contained within $\hat{O}_a$, equating to > 80 % of the total number of samples. We randomly sample 10 sets of $\hat{O}_a$ and $\hat{O}_b$ for each combination of training clusters in order to avoid any bias associated with individual groups of $\hat{O}_a$ or $\hat{O}_b$. Due to the natural variability of lithological class distributions, T_a sampled from low numbers of T_a regions often did not contain representatives for all 13 lithological classes. This leads to biased accuracy

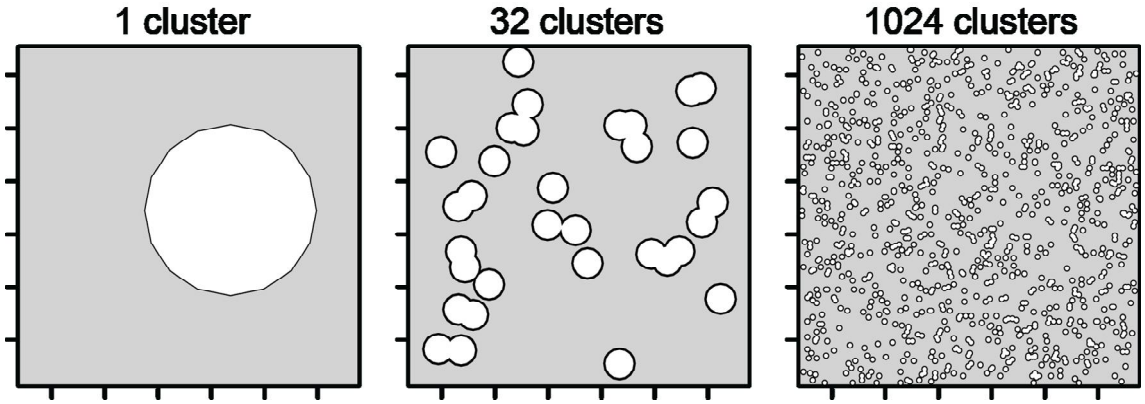

Figure 4.2 Example of T_a spatial distributions for 1, 32 and 1024 clusters. Grey areas outside T_a regions were used to obtain T_b for MLA accuracy and kappa comparisons.

assessments due to the incorporation of samples that cannot be accurately classified. Therefore, those classes not represented in $\hat{O}_a$ were eliminated from their corresponding $\hat{O}_b$ when evaluating MLA predictions.

Airborne geophysical data used in this study contained a Digital Elevation Model (DEM, ASL), Total Magnetic Intensity (TMI, nT) and four Gamma-Ray Spectrometry (GRS) channels comprising Potassium (K %), Thorium (Th ppm), Uranium (U ppm) and Total Count channels. The Landsat 7 ETM+ data contained eight bands supplied with Level 1 processing applied. Landsat 7 ETM+ band 8, which covers the spectral bandwidths of bands 2, 3 and 4 (Williams 2009) was not included.

4.3. Methods

4.3.1. Pre-processing

Geophysical data were transformed to a common projected coordinate system WGS84 UTM zone 54S, using bilinear interpolation. All inputs were resampled to a common extent (12.8 km x 12.8 km) and pixel resolution (50 m x 50 m), resulting in image dimensions of 256 x 256 pixels (65,536 samples). To enhance their relevance to the task of lithology discrimination, projected data were processed in a variety of ways specific to the geophysical property they represent. An account of pre-processing steps implemented to generate input data is provided in Appendix C (C.1 – Data*)*. Spatial coordinates, Eastings (m) and Northings (m), obtained from the location of pixel centres were included resulting in a total of 27 variables available for input. Processed input data were standardised to zero mean and unit variance. Highly correlated data, with mean Pearson's correlation coefficients > 0.8 associated with a large proportion of other data, were eliminated resulting in a total of 17 inputs available for MLA training and prediction. Relative normalised variable importance was calculated by generating Receiver Operating Curves (ROC, Provost & Fawcett 1997) for all pair-wise class combinations and obtaining the area under the curve for each pair. The maximum area under the curve across all pair-wise comparisons was used to define the importance of variables with respect to individual classes (Kuhn *et al.* 2012).

Table 4.2 MLA specific parameters evaluated during classifier training. Note RF parameters presented indicate those used for all input variables (All Data), when inputting only spatial coordinates (XY Only) there is only one possible mtry value (2).

MLA	Parameter	1	2	3	4	5	6	7	8	9	10
NB	*usekernel*	FALSE	TRUE	-	-	-	-	-	-	-	-
kNN	*k*	1	3	5	7	9	11	13	15	17	19
RF	*mtry*	2	3	5	6	8	9	11	12	14	16
SVM	*C*	0.25	0.5	1	2	4	8	16	32	64	128
ANN	*size*	5	8	11	13	16	19	22	24	27	30

4.3.2. Classification model training

Table 4.2 indicates the MLA parameter values assessed in this study. Optimal parameters were selected based on maximum mean accuracies resulting from 10-fold cross-validation. MLA classification models were trained using selected parameters on the entire set of samples in $\hat{O}_a$ prior to prediction evaluation. Information on the R packages and functions used to train MLA classification models and details regarding associated parameters are provided in Appendix C (C.2 – MLA software and parameters).

4.3.3. Prediction evaluation

Overall accuracy and Cohen's kappa statistic (Cohen 1960) are commonly used to evaluate classifier performance (Lu & Weng 2007). Overall accuracy treats predictions as either correct or incorrect and is defined as the number of correctly classified T_b divided by the total number of T_b. The kappa statistic is a measure of the similarity between predictions and observations while correcting for agreement that occurs by chance (Congalton & Green 1998). We do not use the area under ROC to evaluate MLA predictions because multiclass ROC becomes increasing intractable with a large number of classes (> 8, Landgrebe & Paclik 2010). We visualise the spatial distribution of prediction error and assess their geological validity by plotting MLA predictions in the spatial domain and by comparing the locations of misclassified samples.

4.4. Results

In this section we present the results of our comparison of the five MLAs trialled in this study. Initially, we assess the effect of changes in the spatial distribution of training data on the relative importance of input variables and MLA parameter selection. This is followed

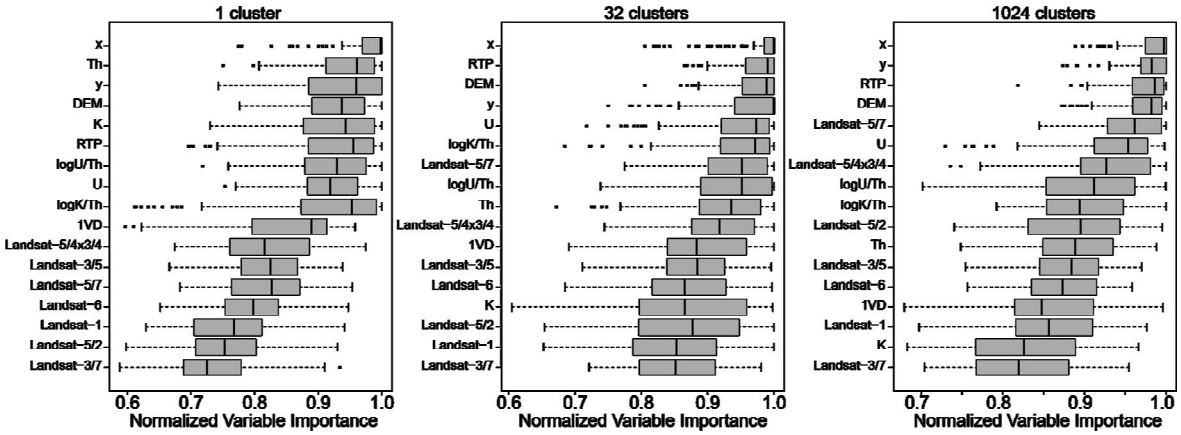

Figure 4.3 Mean ranked normalised variable importance for 1, 32 and 1024 $\hat{Q}_a$ clusters using all data after the removal of highly correlated variables. Boxplots indicate the distribution of combined class univariate importance across all 10 sets of T_a.

by a comparison of MLA T_b statistics in light of variations in the spatial distribution of T_a and the inclusion of explicit spatial information.

Figure 4.3 indicates variations in the importance of input variables with respect to changes in the spatial distribution of T_a. A large range of importance suggests that the usefulness of a particular input at discriminating individual classes is highly variable. Top ranked variables with relatively narrow distributions include X and Y spatial coordinates, DEM and Reduced-to-Pole (RTP) TMI. Less important inputs represent GRS channels and ratios with Landsat ETM+ band 5 variables. Lowest ranked variables are 1st Vertical Derivative (1VD) of RTP TMI and the remaining Landsat ETM+ bands. As the number of T_a regions increases, GRS variables decrease in importance while ratios with Landsat ETM+ 5 increase in importance. The X coordinate is consistently the most important variable, reflecting the presence of approximately north-south trending geological structures.

Figure 4.4 compares MLA cross-validation accuracy with respect to classification model parameter and number of T_a clusters. NB, SVM and ANN cross-validation accuracies are more sensitive to model parameters than to the number of T_a clusters. In contrast, RF and kNN cross-validation accuracies are relatively stable. Fluctuations in cross-validation accuracy are observed for all MLAs with respect to variations in T_a clusters, indicating the sensitivity of MLA classifiers to individual sets of T_a.

Figure 4.5 provides MLA T_b accuracy comparisons with respect to numbers of T_a clusters across the three input variable combinations. All MLAs generate increasingly accurate predictions and exhibit less variation between different T_a for larger numbers of (smaller)

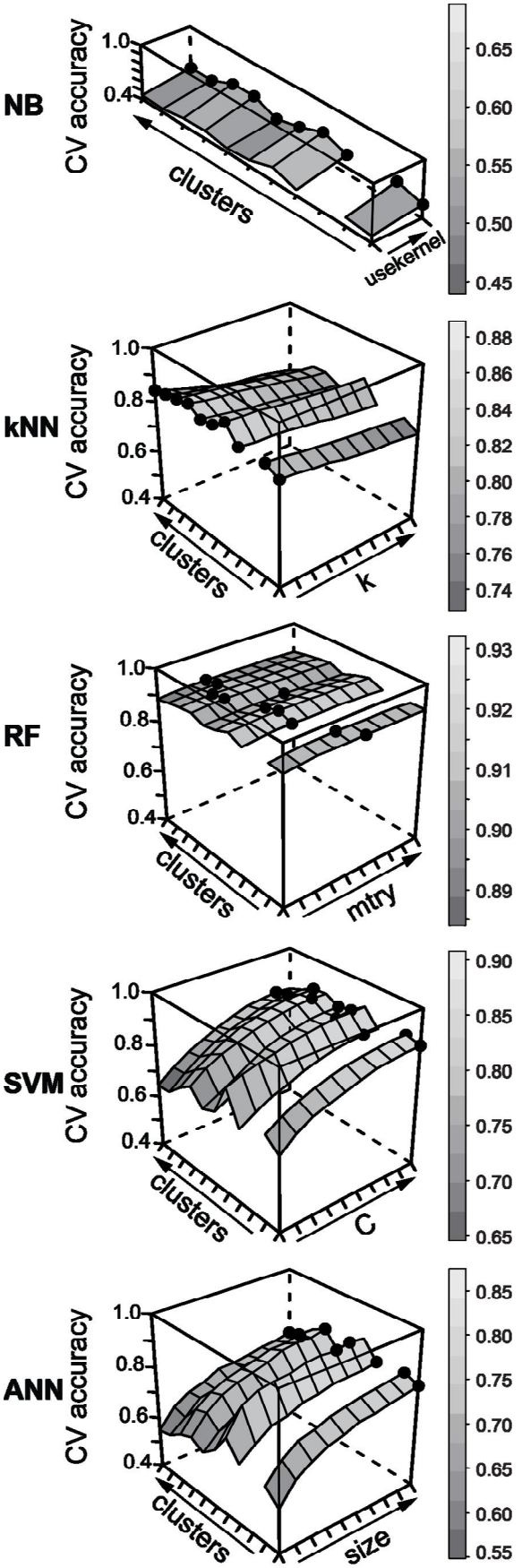

Figure 4.4 Comparison of MLA cross-validation accuracies as a function of classification model parameter and number of T_a clusters. Selected parameter value = black circle. Refer to Table 4.2 for corresponding MLA parameter values.

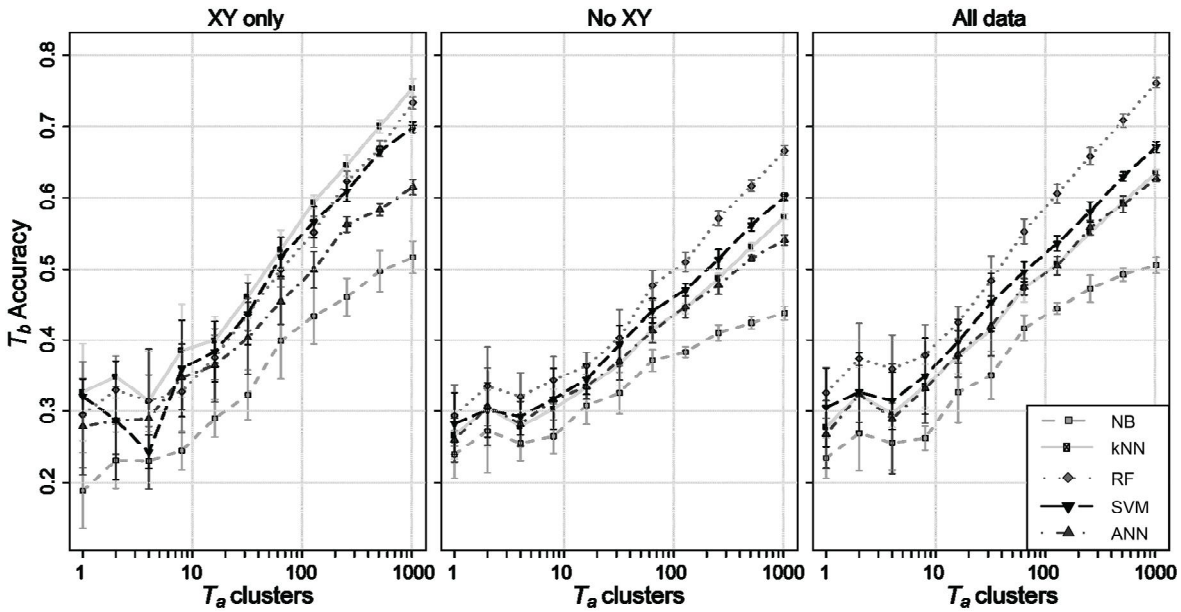

Figure 4.5 Comparison of MLA mean T_b accuracy with respect to variations in the number of T_a clusters. Error bars represent one standard deviation from mean accuracy calculated from a maximum of 10 sets of T_b.

T_a clusters. For 32 T_a clusters and when incorporating spatial coordinates, either exclusively or inclusively, substantial differences between MLA accuracies are not indicated. A < 0.10 increase in T_b accuracy is observed when including spatial coordinates and training MLAs using > 32 T_a clusters, compared to using only geophysical data as input. NB consistently generates the lowest mean T_b accuracies of all MLAs. Where the number of T_a clusters is > 32 and when using only spatial coordinates, kNN, RF and SVM obtain T_b accuracies within 0.05 of each other up to a maximum accuracy of ~ 0.70–0.75 when using 1024 clusters. When utilising only geophysical data and all available data (including spatial coordinates) RF generates substantially higher T_b accuracies than all other MLAs for T_a clusters > 32.

Table 4.3 compares the mean and standard deviation of cross-validation and T_b accuracies and kappa statistics when using all input variables. MLAs exhibit a slight decrease in mean cross-validation accuracy (and kappa) and substantial increase in mean T_b accuracy (and kappa) with increasing numbers of T_a clusters. As the number of T_a clusters increases the difference between mean cross-validation accuracy (and kappa) and mean T_b accuracy (and kappa) decreases, indicating that trained classification models are increasingly over-fitted when presented with smaller numbers of T_a clusters. MLAs were unable to train classification models using the samples provided within T_a or due to cross-validation

Table 4.3 Comparison of MLA cross-validation and T_b accuracy and kappa using all input variables with respect to different numbers of T_a clusters. T_a n indicates the number successfully trained classification models for a given run used to calculate mean and standard deviation. Selected parameters correspond to the MLA parameter (refer to Table 4.2) that obtained the maximum mean cross-validation accuracy and 'count' indicates the number of times this parameter occurred.

			Cross-Validation						T_b			
			Accuracy		Kappa				Accuracy		Kappa	
	T_a	T_a n	Mean	Standard Deviation	Mean	Standard Deviation	Selected parameter	count	Mean	Standard Deviation	Mean	Standard Deviation
NB	1	9	0.618	0.069	0.527	0.073	TRUE	9	0.234	0.028	0.082	0.021
	2	10	0.644	0.030	0.553	0.040	TRUE	10	0.269	0.052	0.122	0.038
	4	9	0.620	0.070	0.534	0.086	TRUE	9	0.255	0.038	0.104	0.032
	8	8	0.615	0.017	0.536	0.024	TRUE	8	0.262	0.017	0.132	0.026
	16	10	0.638	0.059	0.569	0.071	TRUE	10	0.327	0.043	0.196	0.053
	32	10	0.632	0.029	0.568	0.037	TRUE	10	0.351	0.034	0.234	0.043
	64	8	0.630	0.052	0.568	0.056	TRUE	8	0.417	0.018	0.319	0.020
	128	9	0.609	0.021	0.548	0.025	TRUE	9	0.445	0.008	0.355	0.010
	256	10	0.582	0.029	0.516	0.033	TRUE	10	0.473	0.019	0.389	0.020
	512	10	0.573	0.018	0.506	0.021	TRUE	10	0.492	0.009	0.413	0.010
	1024	10	0.543	0.009	0.472	0.010	TRUE	10	0.506	0.012	0.430	0.013
kNN	1	9	0.832	0.024	0.787	0.023	1	9	0.278	0.047	0.120	0.053
	2	10	0.852	0.019	0.806	0.020	1	9	0.323	0.038	0.166	0.031
	4	9	0.850	0.007	0.808	0.014	1	8	0.297	0.019	0.153	0.032
	8	9	0.848	0.014	0.808	0.014	1	8	0.336	0.030	0.203	0.033
	16	10	0.861	0.016	0.828	0.022	1	8	0.377	0.024	0.256	0.029
	32	10	0.865	0.009	0.837	0.011	1	9	0.415	0.039	0.306	0.047
	64	8	0.866	0.013	0.840	0.014	1	7	0.474	0.020	0.378	0.022
	128	9	0.867	0.005	0.843	0.005	1	9	0.507	0.006	0.420	0.007
	256	10	0.860	0.007	0.834	0.009	1	10	0.551	0.005	0.472	0.005
	512	10	0.860	0.006	0.835	0.007	1	10	0.594	0.006	0.522	0.007
	1024	10	0.835	0.008	0.806	0.010	1	10	0.635	0.004	0.571	0.004
RF	1	9	0.914	0.010	0.890	0.009	9	4	0.326	0.035	0.148	0.050
	2	10	0.921	0.009	0.896	0.010	9	3	0.374	0.050	0.220	0.049
	4	9	0.924	0.004	0.902	0.006	6	4	0.359	0.048	0.216	0.054
	8	9	0.915	0.007	0.893	0.008	6	3	0.379	0.043	0.242	0.048
	16	10	0.918	0.011	0.899	0.013	5	4	0.425	0.022	0.300	0.028
	32	10	0.918	0.002	0.901	0.004	6	4	0.484	0.034	0.381	0.041
	64	8	0.917	0.007	0.901	0.009	5	2	0.553	0.017	0.466	0.020
	128	9	0.915	0.006	0.900	0.008	5	3	0.606	0.013	0.531	0.016
	256	10	0.910	0.004	0.893	0.005	6	3	0.658	0.012	0.594	0.014
	512	10	0.907	0.003	0.891	0.003	8	4	0.709	0.009	0.654	0.011
	1024	10	0.892	0.005	0.872	0.006	9	3	0.762	0.007	0.718	0.008
SVM	1	9	0.877	0.019	0.843	0.020	64	6	0.305	0.055	0.144	0.059
	2	6	0.890	0.011	0.857	0.012	128	4	0.327	0.055	0.175	0.063
	4	7	0.891	0.006	0.862	0.007	128	5	0.315	0.040	0.184	0.044
	8	7	0.876	0.015	0.846	0.017	128	4	0.350	0.054	0.217	0.057
	16	9	0.888	0.011	0.862	0.016	64	5	0.398	0.031	0.280	0.037
	32	7	0.882	0.008	0.857	0.009	128	5	0.453	0.041	0.351	0.046
	64	8	0.884	0.012	0.860	0.014	64	5	0.497	0.014	0.404	0.015
	128	7	0.877	0.008	0.855	0.010	64	4	0.536	0.011	0.453	0.014
	256	8	0.870	0.006	0.846	0.008	128	6	0.580	0.014	0.505	0.017
	512	10	0.861	0.004	0.836	0.004	128	5	0.630	0.006	0.563	0.007
	1024	10	0.832	0.007	0.802	0.009	64	7	0.671	0.008	0.613	0.009
ANN	1	9	0.816	0.036	0.765	0.039	30	8	0.267	0.048	0.121	0.056
	2	10	0.838	0.019	0.788	0.022	30	7	0.324	0.040	0.180	0.027
	4	9	0.840	0.024	0.795	0.033	30	5	0.289	0.077	0.172	0.064
	8	9	0.828	0.015	0.783	0.015	30	8	0.332	0.049	0.208	0.047
	16	10	0.837	0.021	0.798	0.028	30	6	0.381	0.033	0.267	0.035
	32	10	0.826	0.015	0.790	0.018	30	8	0.420	0.042	0.314	0.044
	64	8	0.824	0.026	0.789	0.030	30	8	0.475	0.012	0.379	0.014
	128	9	0.808	0.012	0.772	0.015	30	8	0.505	0.013	0.418	0.017
	256	10	0.787	0.012	0.747	0.015	30	8	0.558	0.010	0.479	0.011
	512	10	0.765	0.012	0.723	0.013	30	8	0.591	0.012	0.517	0.013
	1024	10	0.735	0.008	0.687	0.009	30	6	0.627	0.005	0.561	0.005

partitions not containing at least one sample from each class. SVM was more likely to encounter errors during training than other MLAs.

Figure 4.6, Figure 4.7 and Figure 4.8 compare the spatial distribution of MLA lithology predictions and location of misclassifications for selected numbers of T_a clusters. Where inputs consist only of spatial coordinates (Figure 4.6), T_a must be distributed across the entire study area in order to train MLAs that generate predictions closely mimicking the spatial distribution of lithologies. Despite lower overall accuracies when provided with only geophysical data (Figure 4.7), all MLAs generate lithological predictions that indicate major geological structures and trends, i.e. approximate locations of contacts between lithologies. However, the correct lithologies are not always identified. Nonetheless, a large amount of high-frequency noise is present in all MLA predictions when spatial coordinates are excluded. Using T_a dispersed across the study area and all available variables (Figure 4.8) not only generates geologically plausible predictions but the degree to which predictions are affected by high-frequency noise is greatly reduced.

Figure 4.9 provides a comparison of the processing time required to train MLAs and generate predictions given input variables and the spatial distribution of T_a. As the number of T_a clusters increases all MLAs incur additional computation cost. Overall, kNN takes consistently less and SVM consistently more time to run than other MLAs. Increases in processing time are observed when more variables are used as input, with NB and RF taking considerably longer to run. ANN processing times do not substantially increase when including additional data. In contrast to other MLAs, NB requires more time to generate predictions than to train classification models.

4.5. Discussion

4.5.1. Machine learning algorithms compared

Of the MLAs trialled in this study, RF performed well across several aspects evaluated, such as stability, ease of use, processing time and prediction accuracy. A strong aspect of RF performance was its relative insensitivity to variations in parameter values, indicating that RF reduces the potential for classifier overfitting (Breiman 2001). The range of RF cross-validation accuracies was less than ~ 0.05 with increasing *mtry* values. In contrast, SVM (*C*) and ANN (*size*) cross-validation accuracies range was > 0.35 across all

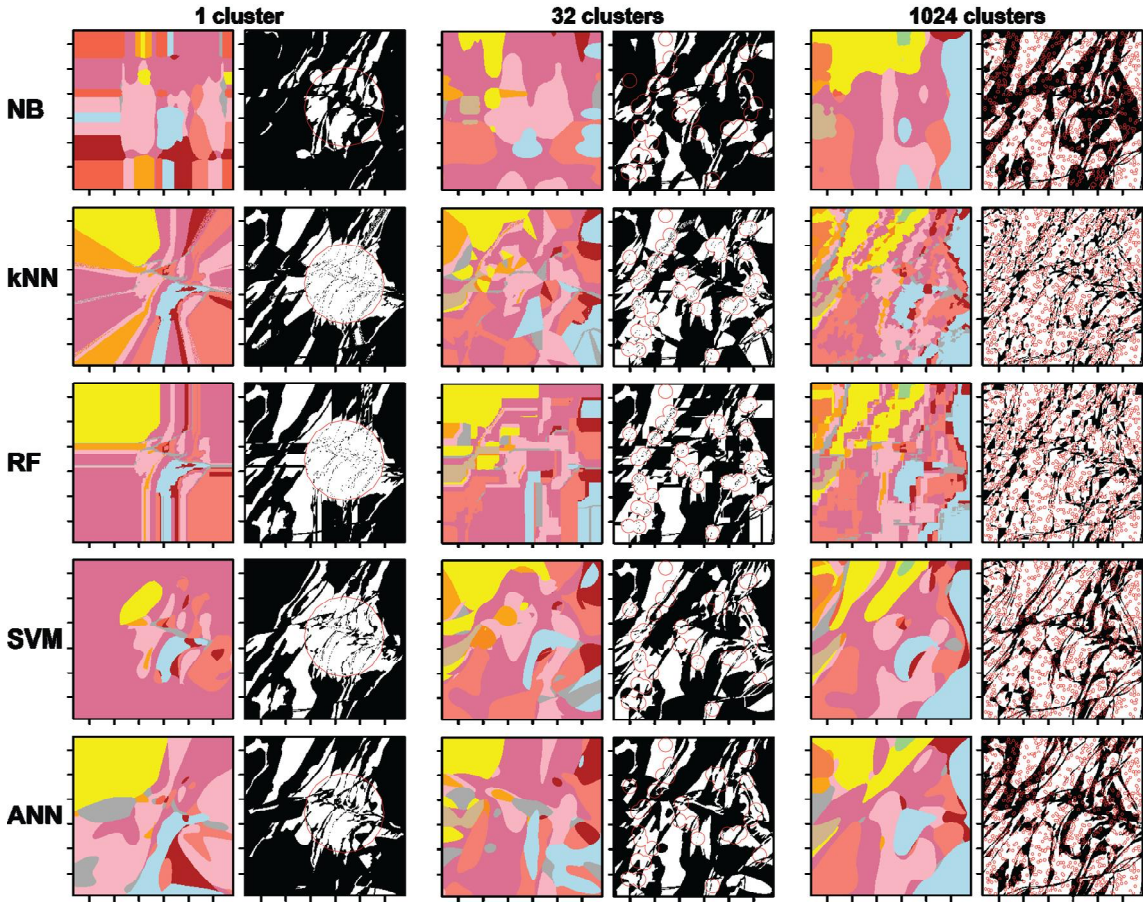

Figure 4.6 Visualisation of the spatial distribution of MLA lithology class predictions (refer to Figure 4.1 for reference map and key to class labels) and misclassified samples (black pixels) for 1, 32 and 1024 T_a clusters (red circles) using X and Y spatial coordinates (XY Only) as inputs.

parameter values. The relative instability of SVM and ANN highlights the need to effectively search a large parameter space to optimise their performance. In this study, we have simplified comparisons by deliberately limiting the search to one parameter for SVM (ó estimated) and ANN (*decay* constant). In situations where multiple parameters require optimisation grid searches can be employed. However, this can be very computationally expensive with fine search grids or, as in the case of coarse search grids, they may not be effective at identifying optimal parameters for a given dataset due to the effect of discretisation (Guyon 2009). Alternatively, a multi-stage nested search could be employed to reduce computational cost and improve parameter selection outcomes (Oommen *et al.* 2008). A nested search initially identifies a small region of parameter space using a coarse grid then searches this region using a finer grid of parameter values.

The least computationally costly MLA to implement was kNN. RF, ANN and NB each had similar processing times (when using geophysical variables) and took less time to run than

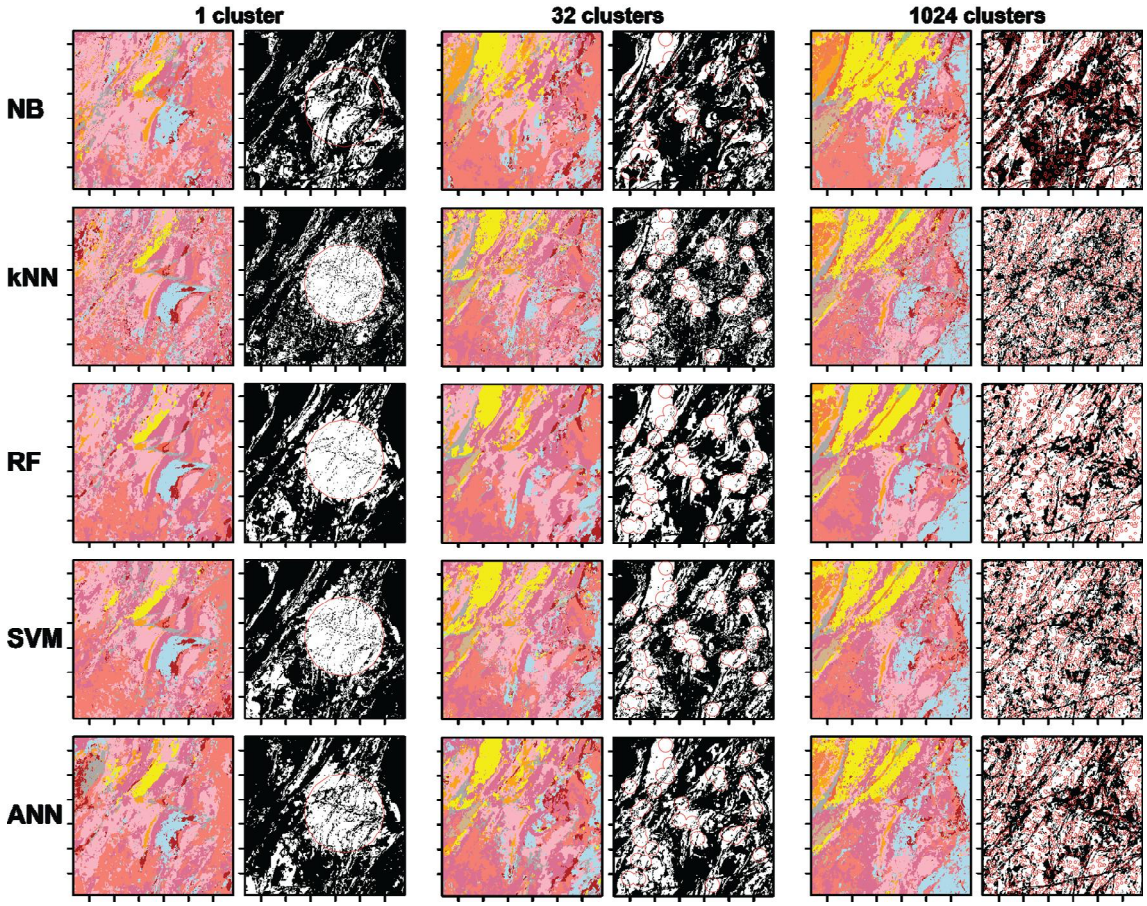

Figure 4.7 Visualisation of the spatial distribution of MLA lithology class predictions (refer to Figure 4.1 for reference map and key to class labels) and misclassified samples (black pixels) for 1, 32 and 1024 T_a clusters (red circles) using geophysical data (No XY) as inputs.

SVM. Rankings of MLAs based on processing time were maintained regardless of an increase in the number of $\hat{O}_a$ clusters or the number of input variables. The use of spatially dispersed T_a resulted in a large increase in the already lengthy SVM processing times.

$\hat{O}_b$ accuracy and kappa results indicate that RF generated substantially higher (between ~ 0.05–0.1) $\hat{O}_b$ mean accuracy and kappa than other MLAs when provided with geophysical data represented by spatially dispersed $\hat{O}_a$. SVM, kNN and ANN all obtained $\hat{O}_b$ mean accuracies and kappa statistics within ~ 0.05, while NB obtained significantly lower T_b accuracies compare to other MLAs. SVM was more likely to encounter convergence errors when training classification models as compared to other MLAs. The good performance of RF in these experiments suggests that its approach to learning, which resembles an adaptive kNN strategy (Hastie *et al.* 2009), provides advantages over other MLAs when applied to spatially constrained supervised classification problems. Furthermore, RF is generally insensitive to noisy input data where the number of relevant variables is

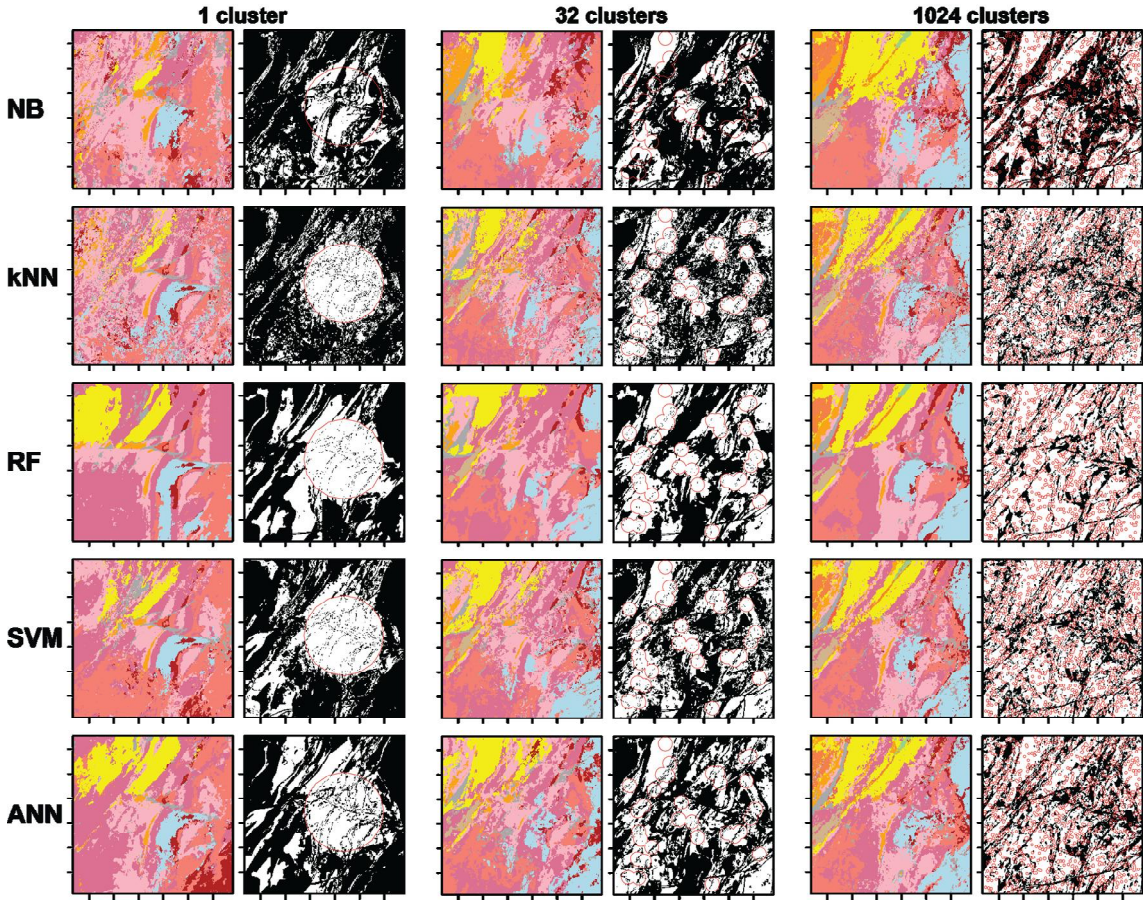

Figure 4.8 Visualisation of the spatial distribution of MLA lithology class predictions (refer to Figure 4.1 for reference map and key to class labels) and misclassified samples (black pixels) for 1, 32 and 1024 T_a clusters (red circles) using spatial coordinates and geophysical data (All Data) as inputs.

comparatively large. In these situations, the effect of variations in the information content of T_a on RF classification models is reduced, improving RF generalisation on unseen data (Breiman 2001; Hastie *et al.* 2009).

4.5.2. Influence of training data spatial distribution

Previous studies into the application of MLAs to remote sensing supervised classification examined the influence of the number of samples in $\hat{O}_a$ on overall predictive accuracy (e.g., Ham *et al.* 2005; Gelfort 2006; Oommen *et al.* 2008; Song *et al.* 2012). These studies found the minimum number of $\hat{O}_a$ samples required to induce accurate MLA classifiers needs to be between 10 % and 25 % of the total number of samples. Furthermore, increasing the number of samples in $\hat{O}_a$ did not lead to improved classification accuracy despite additional computational cost. The experiments conducted in this study clearly indicate that $\hat{O}_a$ spatial clustering is a limiting factor in the efficacy of MLA to predict spatially distributed phenomena while having little effect on processing time.

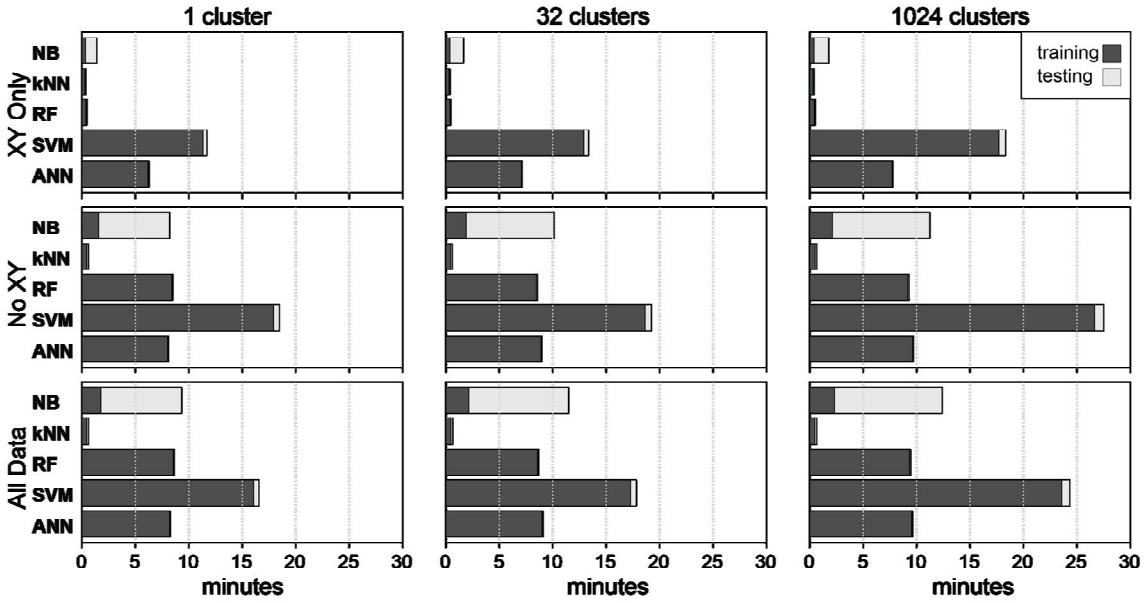

Figure 4.9 Comparison of MLA training (using 10-fold cross-validation) and prediction processing times (minutes). Bars represent the mean time taken to train a MLA classification model and generate predictions for a maximum of 10 sets of T_a and T_b. All processes were executed using a DELL Desktop PC 64-bit Intel Dual Core CPU @ 3.00 GHz and 8 GB of RAM. Cross-validation folds and T_b predictions (divided into groups of 10,000) were run independently using Message Passing Interface parallel processing on 2 cores.

As samples within $\hat{O}_a$ become increasingly dispersed across the region of interest, classifier $\hat{O}_b$ accuracy (and kappa) improves dramatically. Moreover, the difference between cross-validation performance estimates and spatially disjoint $\hat{O}_b$ accuracy decreases, indicating better classifier generalisation capabilities. Improvements in classifier generalisation with highly dispersed $\hat{O}_a$ are coupled with reduced variability of MLA performance on individual sets of $\hat{O}_a$. These findings imply that for real-world mapping applications spatially scattered field observations are most likely to generate stable MLA predictions that accurately reflect the spatial distributions of lithologies. Given appropriately distributed $\hat{O}_a$, the proposed methods are applicable to problems such as: generating rapid first-pass predictions in complex geological terranes; reducing the subjectivity with which geological interpretations are formulated; and validating pre-existing interpreted geological maps.

4.5.3. Using spatially constrained data

Spatial dependencies (spatial autocorrelation), i.e. closer observations are more related than those farther apart, are commonly encountered within spatial data (Anselin 1995; Getis 2010; Lloyd 2011). When incorporating spatial coordinates as input data for training

and prediction, MLAs are provided with explicit information linking location and lithology classes. Spatial values, coupled with randomly distributed $\hat{O}_a$ samples over the entire study area, resulted in a best-case scenario in terms of MLA T_b accuracy. However, geologically plausible predictions were only achieved using input variables containing information that reflected the physical properties of the lithologies under investigation. This observation suggests that standard measures of classifier performance, such as accuracy and kappa, do not necessarily provide an indication of the validity of MLA classifications in the spatial domain. Therefore, the use of spatial coordinates should be approached with caution. If, for instance, T_a were collected only from field observations, these data are likely to be spatially clustered due to outcrop exposure, accessibility and time constraints. Using only spatial coordinates with highly spatially clustered $\hat{O}_a$ samples generates classifiers that incorrectly learn a finite range of coordinate values for a particular class and were unable to generalise well to spatially disjoint regions (Gahegan 2000).

An alternative to the use of spatial coordinates as a means of providing MLAs with spatial context would be to incorporate some measure of local spatial relationships during pre-processing using neighbourhood models (Gahegan 2000). In its simplest form, neighbourhood models generate new variables by deriving statistical measures of the values of proximal observations such as mean or standard deviation. In this way neighbourhood models represent local spatial similarity or dissimilarity (Lloyd 2011). An alternative to pre-processing inputs to include information on the spatial context of observations is to code some notion of spatial dependency into the MLA itself. Li *et al.* (2012) developed and tested a SVM algorithm that directly utilised spatial information to improve classifier accuracy. Their SVM variant takes analysed hold-out T_b results obtained in geographical space and using a local neighbourhood identifies the support vectors that are likely to be misclassified. The SVM hyperplane boundary is then adjusted to include these support vectors. This process proceeds iteratively until the change in prediction accuracy reaches some minimum threshold.

4.6. Conclusions

We compared five machine learning algorithms in terms their performance with respect to a supervised lithology classification problem in a complex metamorphosed geological terrane. These machine learning algorithms: Naïve Bayes, *k*-Nearest Neighbours, Random Forests, Support Vector Machines and Artificial Neural Networks, represent the five

general machine learning strategies for automated data-driven inference. Machine learning algorithm comparisons included their sensitivity to variations in spatial distribution of training data and response to the inclusion of explicit spatial information.

We conclude that Random Forests is a good first choice machine learning algorithm for multiclass inference using widely available high-dimensional multisource remotely sensed geophysical variables. Random Forests classification models are, in this case, easy to train, stable across a range of model parameter values, computationally efficient and when faced with spatially dispersed training data, substantially more accurate than other machine learning algorithms. These traits, coupled with the insensitively of Random Forests to noise and overfitting, indicate that it is well suited to remote sensing lithological classification applications.

As the spatial distribution of training data becomes more dispersed across the region under investigation, machine learning algorithm predictive accuracy (and kappa) increases. In addition, the sensitivity of machine learning algorithm categorical predictions to different training datasets decreases. The inclusion of explicit spatial information (i.e. spatial coordinates) proved to generate highly accurate machine learning algorithm predictions when training data was dispersed across the study area. Despite resulting in lower test accuracy (and kappa), the use of geophysical data provided machine learning algorithms with information that characterised geological structural trends. Therefore, combining spatial and geophysical data is beneficial, especially in situations where training data is moderately spatially clustered. Our results illustrate the use of machine learning techniques to address data-driven spatial inference problems and will generalise readily to other geoscience applications.

4.7. Acknowledgements

We thank Malcolm Sambridge and Ron Berry for their constructive comments and discussion which significantly improved the draft manuscript. Airborne geophysics data were sourced from Geoscience Australia and Landsat ETM+ data from the United States Geological Survey. This research was conducted at the Australian Research Council Centre of Excellence in Ore Deposits (CODES) under Project No. P3A3A. M. Cracknell was supported through a University of Tasmania Elite Research Ph.D. Scholarship. We thank the editor and Milos Kovacevic for their constructive comments that have strengthened the scientific content and clarity of presentation.

4.8. Description of supplementary information

Supplementary information is provided in Appendix C. The R programming language, available from The Comprehensive R Archive Network (http://cran.r-project.org/), was used to pre-process data and facilitate statistical and spatial analysis of MLA outputs. There are two Supplementary Information sections covering: (C.1) input data sources and detailed pre-processing steps; and (C.2) R packages and functions used and details regarding MLA parameter options.

CHAPTER 5 – THE UPSIDE OF UNCERTAINTY: IDENTIFICATION OF LITHOLOGY CONTACT ZONES FROM AIRBORNE GEOPHYSICS AND SATELLITE DATA USING RANDOM FORESTS AND SUPPORT VECTOR MACHINES

Published in Geophysics, vol. 78, no. 3, pp. WB113–WB126, 2013.

5.0. Abstract

Inductive machine learning algorithms attempt to recognise patterns in and generalise from empirical data. They provide a practical means of predicting lithology, or other spatially varying physical features, from multi-dimensional geophysical datasets. It is for this reason machine learning approaches are increasing in popularity for geophysical data inference. A key motivation for their use is the ease with which uncertainty measures can be estimated for non-probabilistic algorithms. We compare and evaluate the abilities of two non-probabilistic machine learning algorithms, Random Forests and Support Vector Machines, to recognise ambiguous supervised classification predictions using uncertainty calculated from estimates of class membership probabilities. We formulate a method to establish optimal uncertainty threshold values in order identify and isolate the maximum number of incorrect predictions while preserving the majority of correct classifications. This is illustrated using a case example of the supervised classification of surface lithologies in a folded, structurally complex, metamorphic terrane. We show that: (1) the use of optimal uncertainty thresholds significantly improves overall classification accuracy of Random Forests predictions, but not those of Support Vector Machines, by eliminating the maximum number of incorrectly classified samples while preserving the maximum number of correctly classified samples; (2) Random Forests, unlike Support Vector Machines, was able to exploit dependencies and structures contained within spatially varying input data; and (3) high Random Forests prediction uncertainty is spatially coincident with transitions in lithology and associated contact zones and regions of intense deformation. Uncertainty has its upside in the identification of areas of key geological interest and has wide

application across the geosciences where transition zones are important classes in their own right. The techniques describe in this work are of practical value in prioritising subsequent geological field activities which, with the aid of this analysis, may be focused on key lithology contacts and problematic localities.

5.1. Introduction

Data inference in geophysics is any procedure whereby observed (measured) values are used to infer the spatial distribution of some property, often in the form of model parameters, which is difficult to observe directly. A number of strategies are available to address the data inference problem: deterministic strategies (e.g., Aster *et al.* 2005 and references therein), strategies which make use of ensembles of multiple models (e.g., Sen & Stoffa 1992; Sambridge 1999b; Sambridge 1999a), Bayesian strategies which sample the model parameter space (e.g., Mosegaard & Tarantola 1995; Denison *et al.* 2002; Malinverno 2002; Bodin *et al.* 2012) and supervised machine learning strategies which utilise data-driven (inductive) approaches to generalise from empirical data (e.g., Burl *et al.* 1998; Gahegan 2000; Witten & Frank 2005; Kanevski *et al.* 2009; Marsland 2009). Uncertainty is handled differently by these contrasting data inference strategies. Deterministic methods generally produce a single best data-fitting model with some parameter trade-off and uncertainty estimates provided through the output covariance and resolution matrices (e.g., Spencer & Gubbins 1980), however, the uncertainty estimate is commonly disregarded in the presentation of the results. In contrast, model ensemble, Bayesian and machine learning strategies all embody uncertainty as an output of the modelling approach. In the case of probabilistic models, uncertainty is an inherent output, whereas for non-probabilistic models it is an estimate based on the structure of the inference model (Kotsiantis 2007; Joshi *et al.* 2009; Kanevski *et al.* 2009).

In exploration geophysics, the model parameters being sought often relate to the extent of particular lithologies. In this application, the available data take the form of multiple remotely sensed datasets, comprising airborne geophysics (magnetics, radiometrics and elevation) and satellite multi-spectral data. The inference task is therefore characterised by a high-dimensional input space and a complex and variable set of relationships between those data. A data-driven machine learning approach is a good choice of inference technique for problems of this kind, evident in the growing number of geological remote sensing data inference examples use machine learning (e.g., Oommen *et al.* 2008;

Kovacevic *et al.* 2009; Waske *et al.* 2009). In these studies, the focus is on comparing the categorical lithology predictions generated by machine learning algorithms (MLAs) such as Random Forests (RF) or Support Vector Machines (SVM) with more traditional classifiers, e.g. Maximum Likelihood Classifier, or other machine learning strategies, e.g. Artificial Neural Networks. These studies use only satellite multispectral and hyperspectral data as input variables. Investigations into the use of machine learning prediction uncertainty in conjunction with geophysical data inference are yet to be fully realised. Three strengths of MLAs are their abilities to: (1) utilise information within disparate multi-dimensional datasets that do not have obvious physical connections (Kanevski *et al.* 2009); (2) view the inference task as a combination of correlated subtasks that share information to generate a global model (Ya *et al.* 2007); and (3) with the use of best-practice approaches for the appraisal of machine learning outputs, robust measures of uncertainty are a natural result of the inference process.

We quantify and compare supervised classification prediction uncertainty generated by two popular MLAs, RF (Breiman 2001) and SVM (Vapnik 1995; Vapnik 1998). These algorithms are shown to perform well on a wide range of remote sensing supervised classification tasks (e.g., Foody & Mathur 2004; Ham *et al.* 2005; Pal 2005; Oommen *et al.* 2008; Ceamanos *et al.* 2009; Waske *et al.* 2009; Kovacevic *et al.* 2010; Waske *et al.* 2010; Duro *et al.* 2012). We apply these two algorithms to a challenging supervised classification problem, namely the prediction of surface lithologies in a folded and metamorphosed Palaeoproterozoic terrane using indirect multisource geophysical remote sensing observations collected from airborne and spaceborne platforms. For this demonstration case, we use a reliable, pre-existing geologic map against which we can evaluate predictions and uncertainty results. In the practical, routine, use of the method that we describe, a geologic map of the region under investigation would not yet exist.

Predictions are generated by training classification models using available labelled data (T), divided into training data (T_a) and test data (T_b), that maps input variables to predefined classes. Uncertainty represents a measure of confidence obtained for individual samples in light of possible alternative categorical predictions. The uncertainty relating to individual samples is obtained by transforming a vector of class membership probabilities into a scalar value. We estimate uncertainty using *Variance* (Kohavi & Wolpert 1996), which provides an indication of the distribution of class membership probabilities.

We compare the abilities of RF and SVM to probabilistically identify incorrect categorical predictions. Furthermore, we assess spatial relationships between prediction uncertainties, regions requiring more data and spatially structured geological features such as contact zones. In applied geosciences there are many situations where flagging locations requiring more observations and identifying transition or contact zones in remote or inaccessible areas is advantageous. For example, in exploration geophysics, prioritising fieldwork to collect the maximum amount of high value observations will increase efficiency and reduce operational costs. In addition, particular types of ore deposits such as gold-bearing quartz-veins are spatially associated with shear zones and metamorphosed contacts (Kusky & Ramadan 2002; Jackson 2005). Methods that can highlight zones such as these would be invaluable for mapping and targeting mineralisation. Other potential disciplines that require knowledge of lithology transition zones include hydrogeology and environmental hazard susceptibility mapping, where geological structures and lithological contacts are used as input into the modelling process (Draskovits & Laszlo 2005; Ramli *et al.* 2010).

Recent studies investigating probabilistic predictive geological mapping and the identification of contact zones have used edge detection, geostatistical and data fusion methods to define regions of interest and model abrupt changes in lithology. For example, Taye (2011) used the Laplacian edge enhancement and detection algorithm and more advanced rotation variant template matching method for contact zone identification from integrated geoscience data. Geostatistical conditional simulation of random sets of categorical variables was used by Emery & Gonzalez (2007) to model uncertainties in transition zones between three lithologies. Their models were used to improve Cu grade estimates near lithology boundaries. In their study, spatial continuity between units was modelled via the estimation of geostatistical parameters from indicator variograms. In another study, Slavinski *et al.* (2010) integrated geoscience data for the manual interpretation of major geological units and their contacts. In their study, raw and calibrated data were combined with spatial filters, designed to enhance local textural information and pre-existing geological maps to manually generate interpretations regarding characteristics of particular geological units and associated contacts including the location of major faults. They targeted regions for future field observations where discrepancies arose between pre-existing geologic maps and their interpretations. In contrast, the work developed in this study attempts to objectify methods for inferring geologic units from integrated geophysical data and the identification of misclassified

samples in order to improve classification accuracies. We then use probabilistic methods to highlight locations of potentially significant geological features (transition zones such as faults and irregular contacts) and differences between the predicted and interpreted geologic maps that aids planning future data collection priorities.

5.1.1. The lithology prediction problem

Categorical data are represented by a single variable that has a finite number of discrete, labelled classes whose value is unknown across the extent of a mapped region. We predict discrete lithologies and identify their contact zones using a raster model whose class labels indicate the category covering the centre of a cell/pixel within a regular array. The use of a raster model for assessing continuous variations in uncertainty is easier to conduct than for models based on polygons (Goodchild *et al.* 1992).

The task of predicting discrete classes from multiple input layers can be defined as $y = f(\mathbf{x})$ such that target categories y_c are separable based on the information contained within $\mathbf{x}$ via the function f. Input layers are represented by a d dimensional vector $\langle \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d \rangle$ and classes by c categories $\{y_1, y_2, \dots, y_c\}$. Supervised learning attempts to construct an optimal function f' that maps observables $\mathbf{x}$ to unknowns $\hat{y}$ using a limited set of labelled T_a such that $\hat{y} = f(\mathbf{x})$, where $f' \quad f$, (Gahegan 2000; Hastie *et al.* 2009; Kovacevic *et al.* 2009). Misclassification costs are seldom uniform in real-world data where missing one or more rare or important classes (False Positives) is less desirable than missing particular classes altogether (False Negatives, Provost & Fawcett 1997). For example, in exploration geophysics higher risks are associated with missing certain geological phenomena, such as ore deposit mineralisation, than erroneous predictions in cases where they do not occur. Prediction uncertainty is one method by which we can quantify misclassification cost, reflecting the risk associated with misclassifying a given instance in light of potential alternatives.

Many MLAs have the ability to generate, in conjunction with discrete class labels, a vector representing class membership probabilities, equal in length to the number of classes. The vector p_c quantifies, for a given sample i, the probability that a class is the correct class and provides an opportunity to realistically assess the likelihood of individual predictions (Witten & Frank 2005). In practice, the class that attains the maximal class membership probability for a given sample is deemed to be the correct category (Goodchild *et al.* 1992); however, this does not provide an indication if other classes were candidates for

selection. Consequently, evaluating the distribution of p_c provides a simple means of quantifying categorical prediction uncertainty. We investigate and interpret the spatial distribution of uncertainty in categorical predictions based on the distribution of p_c obtained for individual samples (pixels). Although, we do not attempt to characterise spatial structure or dependency within the categorical predictions, or in the input variables, as it is beyond the scope of this work. In order to understand the origins of the differences in p_c it is important to gain a conceptual understanding of how RF and SVM discriminate between multiple classes from multi-dimensional input variables and, in turn, generate categorical predictions and class membership probabilities for independent samples.

5.1.2. Random Forests

RF, developed by Breiman (2001), is an ensemble classification scheme that utilises a majority vote for class association based on the results of multiple Decision Trees (DT), known as a forest. Randomness is introduced into the algorithm by randomly subsetting a predefined number of input variables (*mtry*) to split at each node of the DT and by bagging. Bootstrap aggregation, or bagging (Breiman 1996), generates T_a for the induction of a classification model for each tree by sampling with replacement a number of samples equal to the number of instances in T_a. This equates to approximately two-thirds of the instances available for training while the remaining samples are used for testing. Bagging is reported to improve classification predictions as long as they are not stable in the presence of altered T_a (Breiman 1996). An advantage of RF is its relative simplicity and ease of use compared to other MLAs.

The Gini Index is used by RF to determine a "best-split" threshold at each node of a decision tree. This method is based on a calculation of the information purity of the child nodes compared to that of their parent node. The Gini Index is defined as:

$$Gini(t) = \sum_{c=1}^{j} g_c\,(1 \quad g_c), \qquad [5.1]$$

where g_c is the probability or the relative frequency of class *c* at node *j* and is given by:

$$g_c = \frac{n_c}{n}, \qquad [5.2]$$

where n_c is the number of samples belonging to class *c* and *n* is the total number of samples within a particular node. For each candidate split, the threshold *t* that defines maximum reduction in class heterogeneity is selected (Breiman *et al.* 1984; Waske *et al.*

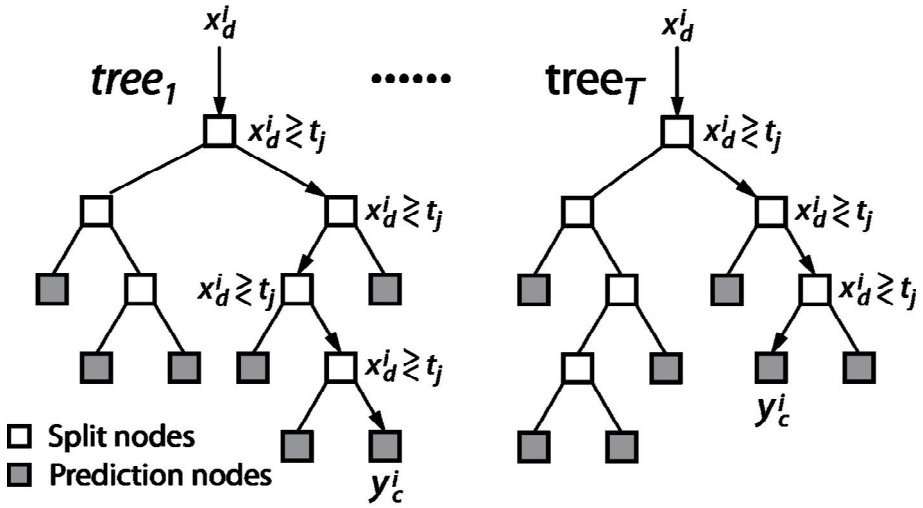

Figure 5.1 Schematic diagram of RF decision tree architecture. $\boldsymbol{x}^i$ *= input data for sample i.* $\boldsymbol{x}^i$ t_j *= split threshold at node j.* y^i *= predicted class for sample i.*

2009). RF grows multiple trees with the results being generally insensitive to noise and model overfitting (Breiman 2001). Predictions are based on a majority vote from the outcome of **T** random decision trees within the forest (Figure 5.1). The probability that a given sample *i* is one of the possible classes *c* is calculated by dividing the frequency of votes for all possible classes by **T** (Liaw & Wiener 2002), such that:

$$p_c = \frac{1}{\mathbf{T}} \sum_{c=1}^{\mathbf{T}} y_c^i, \qquad [5.3]$$

yielding an estimate of class membership probabilities (Hastie *et al.* 2009).

5.1.3. Support Vector Machines

SVM are popular MLAs first described by Vapnik (1995; 1998). It has the ability to define decision boundaries between classes in a high-dimensional input space (Karatzoglou *et al.* 2006; Hsu *et al.* 2010). Basic SVM theory states that for a non-linearly separable dataset containing points from two classes there are an infinite number of hyperplanes that divide the classes. The hyperplane *h* defined as:

$$\boldsymbol{w}\ \mathbf{x} + b = 0, \qquad [5.4]$$

where $\boldsymbol{w}$ and *b* are solved via quadratic optimisation, separates two classes using only a subset of T_a known as support vectors, which defines class marginal hyperplanes. The maximum margin *M* (distance), equal to $\frac{\mathbf{2}}{\|w\|}$, perpendicular to the marginal hyperplanes of

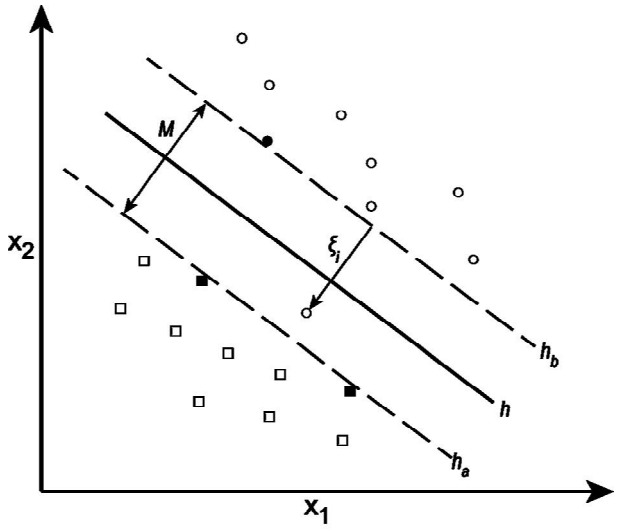

Figure 5.2 Idealised SVM decision boundary for a 2D ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) non-separable linear binary classification problem. Class ***a*** *= and Class* ***b*** *= , shaded symbols indicate support vectors used to calculate the class marginal ($_a$ and $_b$) and maximal marginal (M) decision boundaries. ξ_i = error with regard to support vector misclassification cost.*

the classes in question is taken to represent the optimal decision boundary. Obtaining *M* is equivalent to minimising the objective function:

$$min_{\boldsymbol{w},b} \frac{\|\boldsymbol{w}\|^2}{2}, \qquad [5.5]$$
$$\text{subject to } y_i(\boldsymbol{w}.\mathbf{x}_i + b) \quad 1 , i = 1,2,\dots,c.$$

In non-separable linear cases, SVM find *M* while incorporating a cost parameter *C*, which adjusts the penalty associated with misclassifying support vectors (Figure 5.2). High values of *C* generate more complex prediction functions in order to misclassify as few support vectors as possible (Karatzoglou *et al.* 2006). The objective function must be modified to incorporate this penalty term for wide margined decision boundaries with misclassified T_a:

$$min_{\boldsymbol{w},b,\xi} \frac{\|\boldsymbol{w}\|^2}{2} + C\sum_{i=1}^{N}\xi_i, \qquad [5.6]$$
$$\text{subject to } y_i(\boldsymbol{w}.\mathbf{x_i} + b) \quad 1 \quad \xi_i , i = 1,2,\dots,c.$$

where slack variables ξ_i 0 represent the distance to misclassified support vectors from their respective marginal hyperplanes (Hsu *et al.* 2010).

For non-linear cases, SVM use an implicit transformation of the input data using a kernel function:

$$kern(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle, \quad [5.7]$$

returning the inner product between the positions of pairwise compared input variables ($\mathbf{x}_i$ and $\mathbf{x}_j$) in layer space. The kernel function allows SVM to handle non-linear relationships between classes and variables efficiently and involves projecting samples from **x** space into the kernel space (Hsu *et al.* 2010). SVM deal with binary classification tasks but can be extended to multiclass problems by constructing $\frac{c\,(c-1)}{2}$ binary classification models generating predictions based on a majority vote. This is called the one-against-one method and has been shown to efficiently generate robust results (Hsu & Lin 2002; Kovacevic *et al.* 2010).

Estimation of SVM class membership probabilities is carried out by combining all pairwise probabilities of $\frac{c\,(c-1)}{2}$ binary classification models (Wu *et al.* 2004). In this study we use the Price, Knerr, Personnaz and Dreyfus (PKPD) method (Price *et al.* 1995) for estimating p_c from multiple pairwise binary classifiers. The PKPD method assumes that given the observation **x** and class *y*, the estimated pairwise class probabilities r_{ij} of $\mu_{ij} = \mathrm{P}(y = i | y = i \text{ or } j, \mathbf{x})$ are available. From the i^{th} and j^{th} classes in T_a, a model is obtained which calculates r_{ij} as an approximation of μ_{ij}. An estimate of $p_c = \mathrm{P}(y = c|\mathbf{x})$ for individual classes is obtained using all r_{ij}, such that,

$$p_c = \frac{1}{\sum_{j:j\neq 1} 1/r_{ij} - (c-2)}, \quad [5.8]$$

however as $\sum_{i=1}^{c} p_c \quad 1$, the result must be normalised in order to obtain an estimate of class membership probabilities (Wu *et al.* 2004).

5.2. Data

5.2.1. Tectonic setting and history

Broken Hill is located in the Barrier Ranges of far western New South Wales, Australia. Broken Hill geology features an inlier of the Palaeoproterozoic Willyama Supergroup (WSG, Willis *et al.* 1983). The WSG is a suite of metamorphosed sedimentary, volcanic and intrusive rocks, deposited between 1710–1704 ± 3 Ma (Page *et al.* 2005a; Page *et al.* 2005b). The WSG has a complex outcrop pattern resulting from a long history of folding, shearing, faulting and metamorphism (Stevens 1986; Webster 2004).

Within the sample area defined for this study are 13 lithology classes forming a chronological sequence younging from west to east (Figure 5.3). In general, lithologies comprise a conformable but distinct suite of quartz-feldspathic rocks and composite gneisses, i.e. Thorndale Composite Gneiss and Thackaringa Group consisting of the Alma Gneiss, Alders Tank Formation, Cues Formation and the Himalaya Formation. This suit of rocks is overlain by more psammitic and pelitic rich facies, i.e. Allendale Metasediments and Purnamoota subgroup consisting of the Parnell Formation, Freyers Metasediments and Hores Gneiss overlain by the Sundown Group. The Broken Hill Domain deformation history can be summarised into four events (Stevens 1986; Webster 2004). The first two events of the Olarian Orogeny (1600–1590 Ma, Page *et al.* 2005a; Page *et al.* 2005b) were associated with amphibolite–granulite facies metamorphism. The current northeast-southwest trending regional fabric and outcrop geometry is attributed to the second deformation event. A third event is associated with localised planar or curviplanar Retrograde Schist zones displaying well developed and intense schistosity comprising retrograde mica and chlorotite minerals and strongly deformed metasediment bedding. Retrograde Schist zones generally displace rock units they intersect, fulfilling the role of faults (Stevens 1986). The fourth deformation event, associated with the Delamerian Orogeny (458–520 Ma), is interpreted from gentle dome and basin structures.

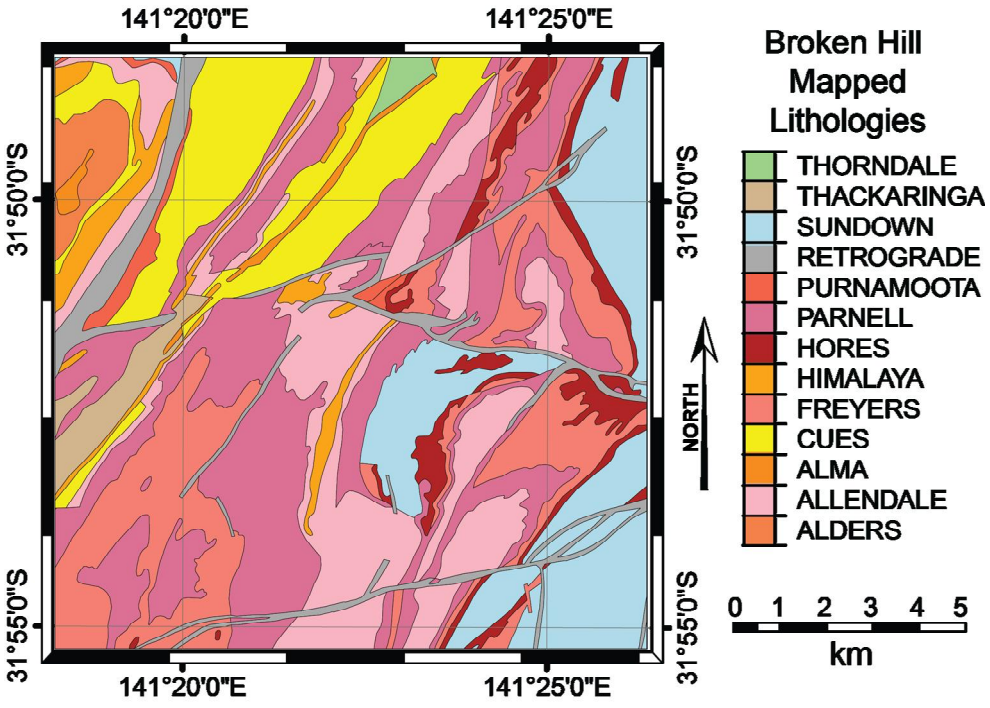

Figure 5.3 Mapped lithologies, after Buckley et al. (2002) within the Broken Hill sample area, western New South Wales, Australia.

5.2.2. Data sources

Airborne geophysical data were sourced from the Australian Geological Survey Organisation (1994), Landsat ETM+ data from the National Aeronautics and Space Administration (2000) and the Broken Hill 1:250,000 digital geologic map was compiled by Buckley *et al.* (2002). Airborne geophysical data were supplied with a Digital Elevation Model (DEM, m ASL), Total Magnetic Intensity (TMI, nT) and three Gamma-Ray Spectrometry (GRS) channels comprising Potassium (K %), Thorium (Th ppm) and Uranium (U ppm). The Landsat 7 ETM+ datasets contain eight bands supplied with Level 1 processing applied. As Landsat 7 ETM+ band 8 covers the spectral bandwidths of bands 2, 3 and 4 (Williams 2009) and is used to sharpen the other bands it was not included as an input dataset in this study.

5.2.3. Data pre-processing

Prior to MLA training and evaluation, input data were projected to WGS84 UTM zone 54S, cropped and resampled using bilinear interpolation to a coincident 50 m pixel resolution with dimensions of 256 × 256 pixels. Several Landsat band ratios were calculated and negative values were replaced in the GRS data prior to smoothing using a 5 × 5 mean focal operator. TMI data was transformed by Reduction to Pole (RTP) and the calculation of the 1st vertical derivative (1VD). Spatial coordinates were calculated from the location of pixel centres. All variables were normalised to zero mean and unit variance. Variables with mean correlation coefficients > 0.8 associated with a large proportion of other data were eliminated in order to reduce the dimensionality of the problem. The removal of highly correlated data resulted in 17 variables available for classification model training and prediction. Table 5.1 provides a description of data sources, pre-processing methods and indicates the selected variables used for machine learning classification model training and evaluation.

5.3. Methods

Data used to train and evaluate MLAs, T, were divided into three components: training and validation data, T_a; and T_b. T_a can be mutually exclusive subsets of the same data, as used in cross-validation (Hastie *et al.* 2009). Cross-validation is a simple but robust method for estimating the performance of trained classification models and for optimising algorithm specific parameters (Kohavi 1995; Witten & Frank 2005). Cross-validation recursively

Table 5.1 Description of input variables: units, resolution; and pre-processing methods.

	Data	Units	Resolution
*	x	Eastings m (integer)	50 m
*	y	Northings m (integer)	50 m
*	DEM	m ASL (float)	~ 21 m
*	RTP	nT (float)	~ 21 m
*	1VD	nT (float)	~ 21 m
	TC	count (float)	~ 21 m
*	K [a b]	% (float)	~ 21 m
*	Th [a b]	ppm (float)	~ 21 m
*	U [a b]	ppm (float)	~ 21 m
*	logK-Th [c b]	ratio (float)	~ 21 m
*	logU-Th [c b]	ratio (float)	~ 21 m
*	Landsat-1-5	DN (8-bit integer)	28.5 m
*	Landsat-6 (sensor 2)	DN (8-bit integer)	57 m
	Landsat-7	DN (8-bit integer)	28.5 m
	Landsat-3-1 [d]	ratio (float)	28.5 m
	Landsat-3-2 [d]	ratio (float)	28.5 m
*	Landsat-3-5 [d]	ratio (float)	28.5 m
*	Landsat-3-7 [d]	ratio (float)	28.5 m
	Landsat-5-1 [d]	ratio (float)	28.5 m
	Landsat-5-2 [d]	ratio (float)	28.5 m
	Landsat-5-4 [d]	ratio (float)	28.5 m
	Landsat-5-7 [d]	ratio (float)	28.5 m
	Landsat-5-4x3-4 [d]	ratio (float)	28.5 m
	Geology [e]	categorical (text)	1: 250,000

* *Indicates variable preserved after dimensionality reduction (see Section 5.2.3 for a description of method employed). Pre-processing codes:* [a] *correct for negative values;* [b] *mean 5 × 5 focal filter;* [c] $\log_e$ *of ratio;* [d] *ratio of Landsat ETM+ bands; and* [e] *rasterised to 50 × 50 m pixel resolution with centre of pixel used as sample label. Note only Landsat band 1 remains after the removal of correlated variables in Landsat ETM+ bands 1-5.*

splits T_a into k mutually exclusive and equally sized subsets, one of which is used as validation data and k 1 are used for training. Summing or averaging over the results of the k folds for a given parameter value provides a performance estimate from which to compare and select optimal parameters. Selected parameters are used to train a classification model using all of the available T_a, including validation data. As cross-validation performance estimates are based on the same data that is used to train the classification model this can result in a misleading performance estimate (Witten & Frank

2005; Hastie *et al.* 2009). Therefore, T_b represents an independent group of samples with known class labels and is used to provide an unbiased evaluation of the performance of the trained classification models.

5.3.1. Training and evaluating algorithms

We have used the R programming language to train and evaluate RF, via *randomForest* (Liaw & Wiener 2002) and SVM, via *kernlab* (Karatzoglou *et al.* 2004) classification models and to estimate and analyse prediction uncertainty. Two R packages in particular, *caret* (Kuhn *et al.* 2012) and *raster* (Hijmans & van Etten 2012) were used as they offer self-contained functions to train and test MLAs and process and analyse gridded spatial data. Appendix D provides more information on the practical implementation of the methods used in this study. Samples representing collocated pixels within the study area were split into equal proportions of T_a and T_b. From T_a spatially randomly sampled data, equal in proportion to 0.01, 0.02, 0.05, 0.10 and 0.25 of the total number of samples in the Broken Hill study area, were selected for machine learning classification model training and parameter selection.

Both RF and SVM require the selection of two parameters to optimise their performance, while SVM also require the selection of a kernel function. For RF these parameters are the number of decision trees **T** to construct and the number of randomly selected input layers (*mtry*) to split at the nodes of individual trees (Liaw & Wiener 2002). **T** was maintained at the default value of 500; whereas 10-fold cross-validation was used to select optimal *mtry* values for the five different T_a sizes. In this case, we used a Gaussian Radial Basis Function (RBF) kernel for SVM, which is a reasonable first choice of kernel for most applications (Hsu *et al.* 2010). SVM require optimisation of support vector misclassification cost, *C* and inverse kernel width, ó. 10-fold cross-validation was used to select optimal *C* values, whereas ó values were estimated using the sigest() function available in *kernlab*. This function estimates 0.1 and 0.9 quantiles of the $||x - x'||^2$ statistics from a sample of T_a as any value for ó between these two quantiles leads to good RBF SVM kernel performance (Karatzoglou *et al.* 2006).

Trained classification models were evaluated using 100 independent randomly sampled T_b each containing 1000 samples. T_b were used to visualise uncertainty with respect to classification error and calculate uncertainty thresholds. In addition, T_b were used to generate confusion matrices and associated recall and precision metrics for individual

classes. Recall, also known as Producer's Accuracy, sensitivity and measures of omission, indicates the probability of reference samples, i.e. a lithology class, being correctly classified. Precision, also known as User's Accuracy, positive predictive value and measures of commission, represents the probability a prediction is correct (Congalton & Green 1998; Witten & Frank 2005). In general, we want a classifier that generates high recall as the cost of misclassifying features of interest, e.g. lithologies, ore deposit mineralisation, is higher than the cost of an erroneous prediction.

5.3.2. Variance

Within the remote sensing literature there are two metrics commonly used to quantify the degree to which the distribution of p_c is concentrated within a particular class (Goodchild *et al.* 1994; Zhu 1997; Brown 1998; van der Wel *et al.* 1998). These metrics represent two methods for calculating uncertainty estimates from p_c. As there is very little difference between the representations of uncertainty from *Variance* or *Entropy* in this situation we use *Variance*.

Variance (Kohavi & Wolpert 1996), similar in form to the quadratic score (Glasziou & Hilden 1989), is given by:

$$Variance = \frac{1 - \sum {p_c}^2}{1 - \sum \left(\frac{1}{c}\right)^2}. \qquad [5.4]$$

Eq. 5.4 is normalised using the lower term in order to generate consistent uncertainty values between 0 and 1 regardless of the number of possible classes.

5.4. Results

Figure 5.4 compares maps of the spatial distribution of *Variance* estimated from class membership probabilities generated by RF and SVM using 0.02, 0.05 and 0.10 proportions of all samples for training. These maps depict (high) uncertainties 0.8 as warm colours and interpreted lithology boundaries are overlaid for reference. There is a decrease in overall uncertainty for both RF and SVM with an increase in T_a sample proportion. Spatially contiguous regions of high uncertainty generated by RF and SVM are located in different regions. When compared with the reference geologic map, it is apparent that there is an increase in the concentration of relatively high RF uncertainty values proximal to lithology contacts. Furthermore, it is the contact zones associated with particular classes

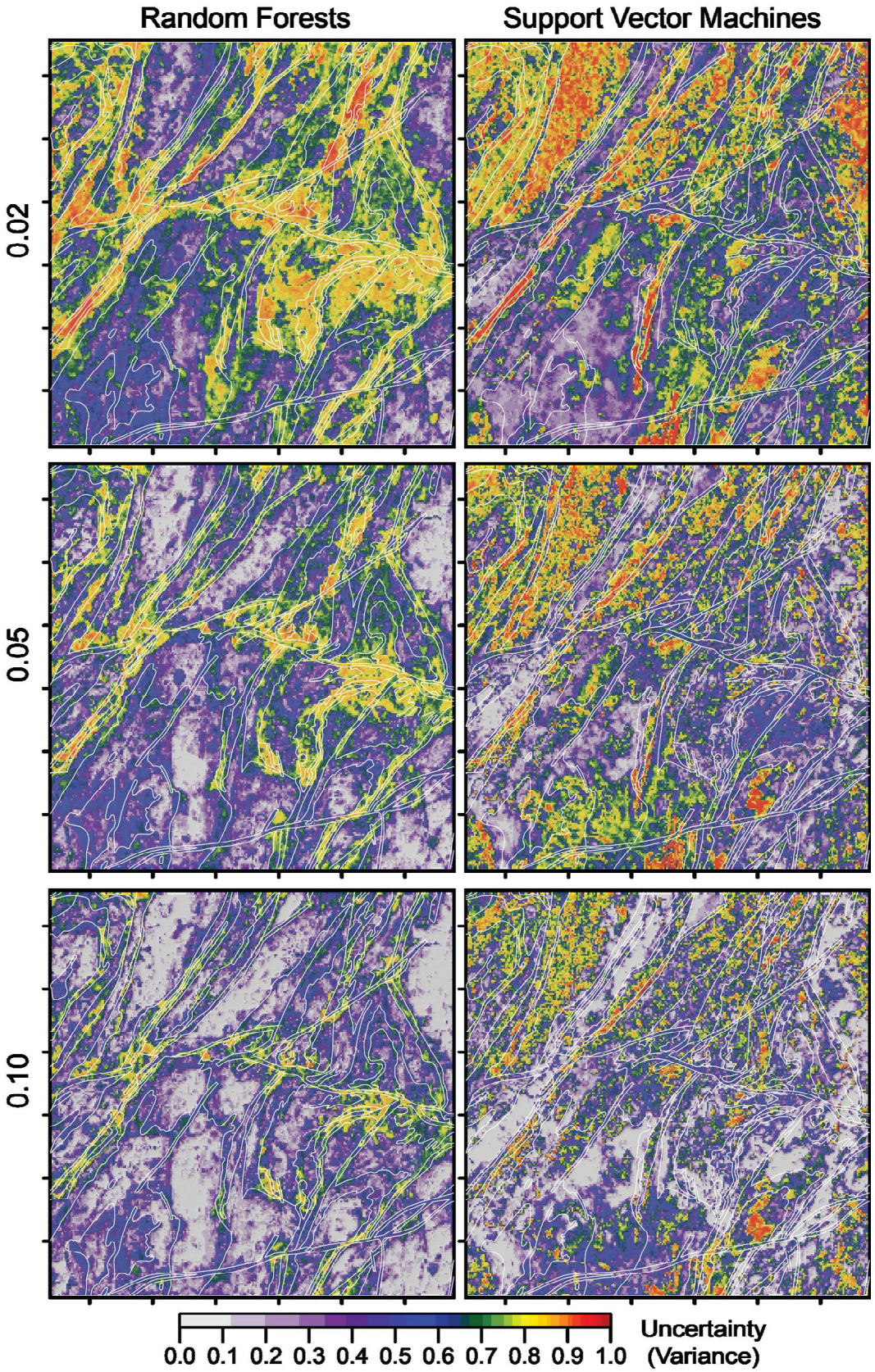

Figure 5.4 Comparison of the spatial distribution of uncertainty estimated from the class membership probabilities generated by RF and SVM using 0.02, 0.05 and 0.10 T_a sample proportions. The axes represent geographical coordinates covering the extent defined in Figure 5.3. White lines correspond to lithology boundaries sourced from Buckley et al. (2002).

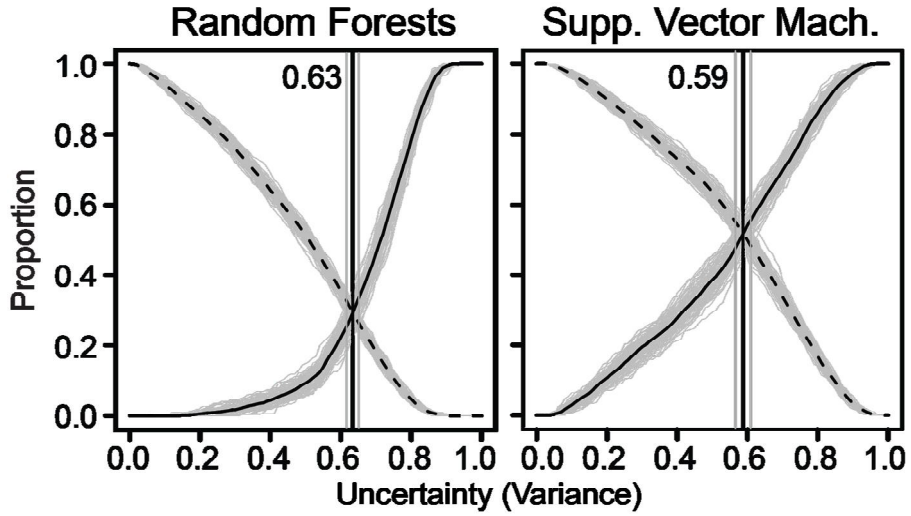

Figure 5.5 T_b error-uncertainty thresholds for 0.05 T_a sample proportion, x-axis represents uncertainty calculated from RF and SVM class membership probabilities and y-axis represents proportion of correctly and incorrectly classified T_b samples. Dashed black curves = mean proportion of correctly classified samples with threshold values greater than a given uncertainty value. Solid black curves = mean proportion of incorrectly classified samples with threshold values less than a given uncertainty value. Grey curves = respective proportions calculated from individual T_b. Black vertical lines = mean curve intersection used to establish the optimal uncertainty threshold values that result in the maximum number of incorrectly classified samples and minimum number of correctly classified samples being identified as unclassified. Grey vertical lines = 2 × standard deviation from mean based on 100 T_b datasets.

such as Alma, Hores and Retrograde classes, or regions of intense deformation resulting in complex geometries, which exhibit the highest RF uncertainties. For SVM, localised regions of high uncertainty are associated with particular units, specifically Allendale, Cues, Freyers, Himalaya, Parnell and Sundown.

Figure 5.5 shows error-uncertainty thresholds for 0.05 T_a sample proportions. Thresholds are calculated from the intersections of curves representing the proportion of correctly classified T_b samples less than a given uncertainty value and the proportion of incorrectly classified T_b samples greater than a given uncertainty value. Samples with uncertainty values including and above thresholds are deemed to be unclassified. Uncertainty thresholds will preserve the maximum amount of correctly classified samples while eliminating the maximum number of incorrectly classified samples only if relatively high uncertainties are associated with misclassified samples. The relative proportion of misclassified samples not identified as unclassified is lower for RF (~ 0.3) than for SVM (~ 0.5). Table 5.2 provides the mean uncertainty thresholds for all T_a sample proportions

calculated from the mean curve intersections resulting from the 100 T_b. SVM is not able to induce a classification model for the 0.01 T_a sample proportion containing ~ 650 samples. In general, RF thresholds are higher than those for SVM, while both algorithms show a decrease in threshold values with increasing T_a sample proportions. For example, the 0.02 T_a proportion mean uncertainty threshold for RF is 0.719 and for SVM it is 0.650. Similarly, the 0.25 T_a proportion thresholds are 0.518 for RF and 0.311 for SVM.

Table 5.3 and Figure 5.6 present T_b overall accuracies and compares these with overall accuracies after unclassified samples have been eliminated based on the uncertainty

Table 5.2 Uncertainty threshold values for RF and SVM across T_a sample proportions. Note SVM was unable to a train classification model using 1 % (~ 650) of the total number of samples.

		T_a sample proportion				
		0.01	**0.02**	**0.05**	**0.1**	**0.25**
RF	mean	0.749	0.719	0.634	0.564	0.518
	2 × st. dev.	0.011	0.013	0.018	0.018	0.030
SVM	mean	-	0.650	0.587	0.480	0.311
	2 × st. dev.	-	0.017	0.022	0.028	0.040

Table 5.3 Comparison of overall T_b accuracies before and after the elimination of unclassified samples using uncertainty thresholds presented in Table 5.2.

			T_a sample proportion				
			0.01	**0.02**	**0.05**	**0.10**	**0.25**
RF	**T_b Accuracy**	Mean Accuracy	0.642	0.709	0.789	0.842	0.893
		95 % Confidence Interval	0.003	0.003	0.002	0.002	0.002
	Uncertainty threshold	Mean Accuracy	0.774	0.835	0.900	0.936	0.970
		95% Confidence Interval	0.004	0.003	0.002	0.002	0.001
		Mean Proportion Unclassified	0.457	0.428	0.381	0.339	0.270
		Change in Overall Accuracy	0.132	0.126	0.111	0.095	0.077
SVM	**T_b Accuracy**	Mean Accuracy	-	0.634	0.715	0.759	0.823
		95 % Confidence Interval	-	0.003	0.002	0.003	0.002
	Uncertainty threshold	Mean Accuracy	-	0.602	0.701	0.739	0.808
		95 % Confidence Interval	-	0.004	0.003	0.003	0.003
		Mean Proportion Unclassified	-	0.490	0.492	0.486	0.485
		Change in Overall Accuracy	-	-0.032	-0.014	-0.020	-0.015

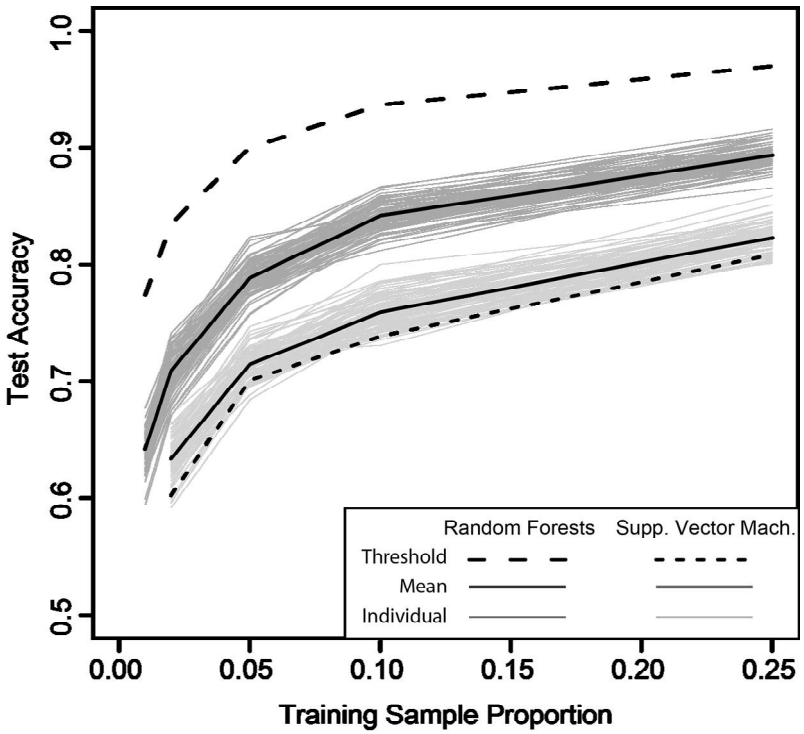

Figure 5.6 Comparison of RF and SVM T_b accuracy (100 groups of 1000 samples) as a function of T_a sample size compared with T_b accuracies generated after uncertainty thresholds applied (see Table 5.2). Exact 95 % Confidence Intervals are less than 0.005 for mean T_b accuracies. Note the significant improvement across all T_a sample sizes after uncertainty thresholds have been applied to RF predictions.

thresholds in Table 5.2. The identification and subsequent removal of unclassified samples has significantly improved RF T_b accuracies for all T_a sample sizes. For example, the resulting mean RF accuracy after removing unclassified samples is 0.774 for 0.01 T_a proportions and 0.97 for 0.25 T_a proportions, equating to 0.132 and 0.077 increases in accuracy respectively. These data indicate that increases in RF accuracy are greater for smaller T_a sample sizes, although there is an equivalent increase in the total number of unclassified samples. In contrast, differences in mean SVM accuracies range from 0.032 for 0.02 T_a proportions to 0.015 for 0.25 T_a proportions indicating a slight accuracy decrease after the isolation of samples deemed to be unclassified. The mean proportion of unclassified SVM predictions consistently represents close to half of the samples in T_b.

Table 5.4 presents the confusion matrices for the T_b predictions generated by RF and SVM classification models using 0.05 T_a sample proportions after isolating unclassified samples. Confusion matrix cell values represent average counts from 100 T_b. Table 5.5 provides a comparison between the percentage of unclassified samples, recall and precision rates and their respective differences for individual classes obtained from the RF and SVM

Table 5.4 RF and SVM confusion matrices for 0.05 Ta sample proportion classification models. Cell values reflect average counts of confusion matrices generated from 100 Tb after isolation of unclassified samples. Note class names are abbreviated from those in Figure 5.3.

RANDOM FORESTS		PREDICTION												
		ALDE.	ALLEN.	ALMA	CUES	FREY.	HIM.	HOR.	PARN.	PURN.	RETRO.	SUND.	THACK.	THORN.
REFERENCE	ALDE.	12.9	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	ALLEN.	0.0	78.0	0.0	0.8	0.0	1.7	0.0	5.9	0.0	0.0	0.0	0.0	0.0
	ALMA	1.3	0.0	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
	CUES	0.2	0.1	0.0	87.5	0.0	0.4	0.0	0.6	0.1	0.1	0.0	0.0	0.0
	FREY.	0.0	0.3	0.0	0.0	80.1	0.0	0.2	8.9	0.0	0.0	0.6	0.0	0.0
	HIM.	0.0	1.5	0.0	1.1	0.0	10.5	0.0	0.4	0.0	0.0	0.0	0.0	0.0
	HOR.	0.0	0.1	0.0	0.0	1.4	0.0	2.8	0.0	0.0	0.0	1.8	0.0	0.0
	PARN.	0.0	3.9	0.0	2.0	10.8	0.0	0.0	154.0	0.0	0.0	0.1	1.0	0.0
	PURN.	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	2.4	0.5	0.0	0.0	0.0
	RETRO.	0.0	3.0	0.0	0.4	2.5	0.0	0.0	3.8	0.4	10.7	4.1	0.0	0.0
	SUND.	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	103.9	0.0	0.0
	THACK.	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.8	0.0	0.0	0.0	10.7	0.0
	THORN.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0

SUPPORT VECTOR MACHINES		PREDICTION												
		ALDE.	ALLEN.	ALMA	CUES	FREY.	HIM.	HOR.	PARN.	PURN.	RETRO.	SUND.	THACK.	THORN.
REFERENCE	ALDE.	7.8	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	ALLEN.	0.0	45.9	0.0	0.5	3.1	1.4	0.4	11.9	0.0	1.0	1.4	0.3	0.0
	ALMA	0.9	0.0	1.1	0.4	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.4
	CUES	0.0	0.5	0.3	18.7	0.0	0.9	0.0	3.1	0.2	0.2	0.0	0.1	0.5
	FREY.	0.0	4.2	0.0	0.0	51.2	0.1	4.3	16.6	0.0	1.3	3.6	0.1	0.0
	HIM.	0.0	2.0	0.0	0.3	0.0	4.1	0.0	1.1	0.0	0.0	0.0	0.7	0.0
	HOR.	0.0	1.0	0.0	0.0	6.9	0.0	11.9	1.4	0.0	0.6	4.4	0.0	0.1
	PARN.	0.0	12.3	0.0	2.3	12.2	0.6	0.6	104.7	0.0	1.5	2.1	2.4	0.0
	PURN.	0.0	0.6	0.0	0.3	0.0	0.1	0.0	0.3	1.2	0.3	0.0	0.0	0.0
	RETRO.	0.4	3.1	0.0	0.3	2.9	0.2	0.6	6.0	0.3	6.1	3.9	0.2	0.0
	SUND.	0.0	2.7	0.0	0.1	3.3	0.0	5.1	1.6	0.0	1.6	72.8	0.0	0.0
	THACK.	0.0	0.1	0.0	0.1	0.2	0.4	0.0	2.1	0.0	0.0	0.0	15.7	0.0
	THORN.	0.0	0.0	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0

Table 5.5.Comparison of RF and SVM class dependent measures of recall and precision rates and their respective differences as obtained from the 0.05 T_a proportion classification models. Recall and precision rates were obtained from the confusion matrices in Table 5.4. Differences are calculated by subtracting results of SVM from those of RF. Class names are abbreviated from those in Figure 5.3.

	RF			SVM			Difference		
	% Unclass.	Recall	Precision	% Unclass.	Recall	Precision	% Unclass.	Recall	Precision
ALDE.	0.315	0.979	0.899	0.573	0.955	0.859	-0.258	0.025	0.039
ALLEN.	0.450	0.903	0.896	0.581	0.698	0.635	-0.131	0.205	0.261
ALMA	0.762	0.122	1.000	0.655	0.391	0.521	0.107	-0.268	0.479
CUES	0.260	0.985	0.944	0.797	0.765	0.804	-0.537	0.220	0.139
FREY.	0.398	0.890	0.846	0.456	0.630	0.642	-0.058	0.260	0.203
HIM.	0.583	0.775	0.837	0.748	0.506	0.531	-0.165	0.270	0.306
HOR.	0.853	0.462	0.921	0.363	0.453	0.522	0.490	0.009	0.399
PARN.	0.310	0.896	0.883	0.443	0.756	0.704	-0.133	0.141	0.179
PURN.	0.708	0.745	0.832	0.757	0.424	0.723	-0.049	0.321	0.109
RETRO.	0.517	0.430	0.948	0.535	0.256	0.480	-0.019	0.174	0.468
SUND.	0.231	0.998	0.941	0.356	0.836	0.826	-0.126	0.162	0.115
THACK.	0.430	0.925	0.909	0.089	0.846	0.806	0.341	0.079	0.103
THORN.	0.416	1.000	0.984	0.060	0.814	0.802	0.356	0.186	0.182

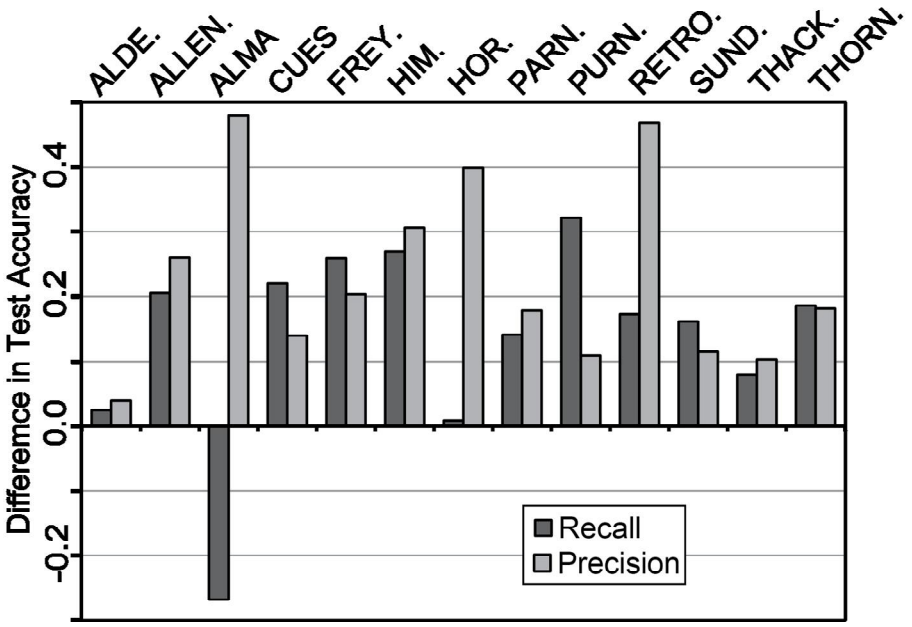

Figure 5.7 Difference between RF and SVM recall and precision rates for the 13 classes in the Broken Hill study area. Class names are abbreviated from those presented in Figure 5.3.

confusion matrices. These Tables indicate the most difficult units for RF to correctly classify, with recall rates of less than or close to 0.5, are the Alma, Hores and Retrograde classes. All precision rates generated by RF are > 0.8 suggesting the majority of its predictions are correct. Interestingly, Alma (0.762) and Hores (0.853) have high percentages of unclassified samples indicating high uncertainties were associated with these classes. In comparison, Alma, Himalaya, Hores, Purnamoota and Retrograde classes are both challenging units for SVM to correctly classify and predict with recall rates close to or less than 0.5. Reflected in unclassified sample percentages, SVM has associated high uncertainties with Cues (0.765), Himalaya (0.748) and Purnamoota (0.757) classes. Very low percentages of unclassified samples (< 0.1) and high recall and precision rates (> 0.8) indicates SVM has accurately predicted Thackaringa and Thorndale samples with confidence. Figure 5.7 shows the differences between recall and precision rates generated by subtracting SVM results from those obtained by RF. RF predictions result in consistently higher probabilities of being correct than those obtained by SVM except in the recall rates for the Alma class. The contrasting difference in recall value for the Alma class relates to its narrow continuous geometry and close association with the dominant Cues formation.

Figure 5.8 depicts predicted lithology maps resulting from RF and SVM (using 0.05 T_a proportions) after unclassified samples have been identified (indicated as black pixels).

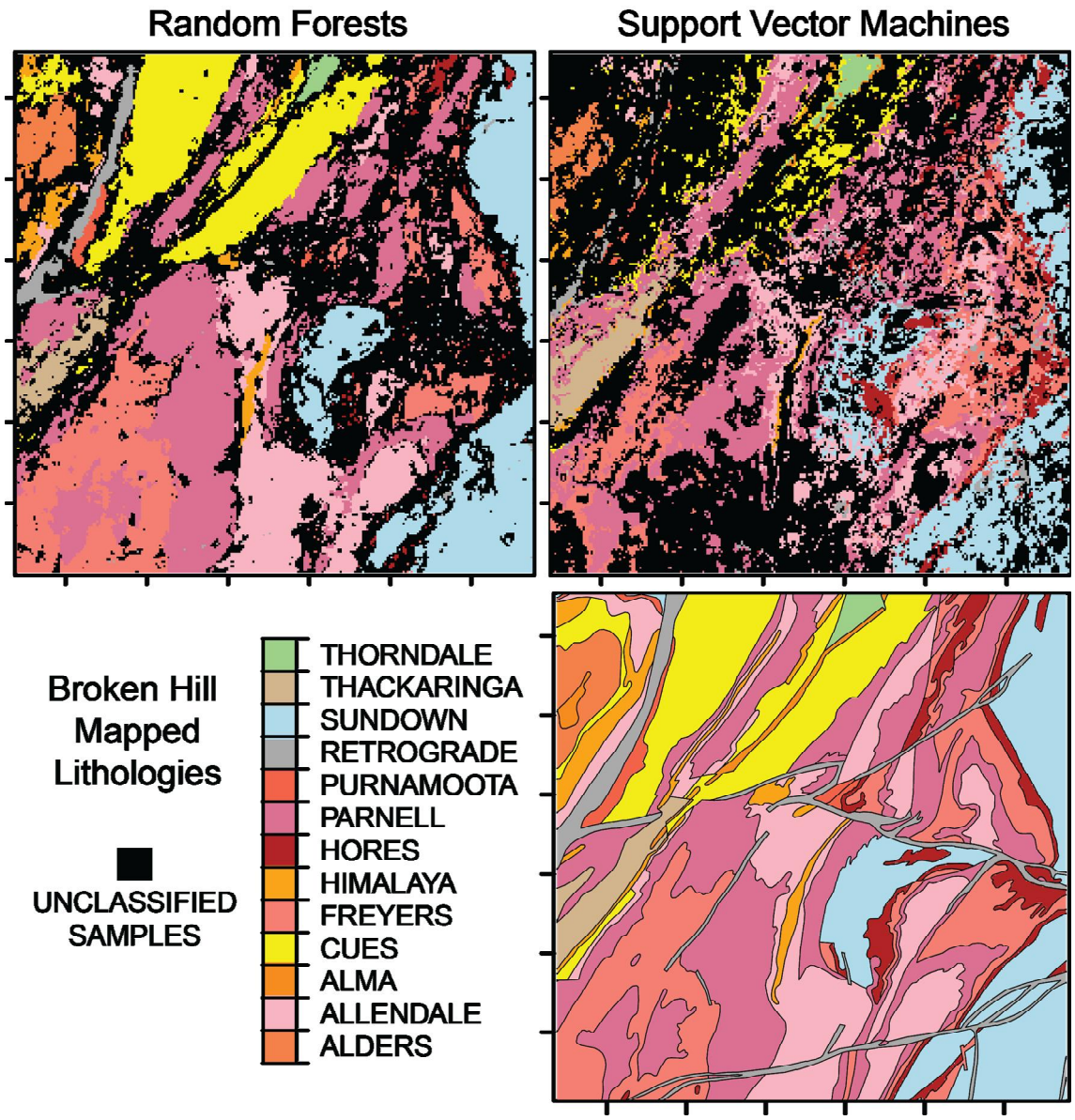

Figure 5.8 Comparison of the spatial distribution of RF and SVM lithology predictions and unclassified samples identified using uncertainty thresholds using 0.05 T_a sample proportions. The interpreted lithology map, after Buckley et al. (2002), is provided for reference. The axes represent geographical coordinates covering the extent defined in Figure 5.3.

The RF predictions associate high uncertainties in the central eastern and western regions of the sample area, coincident with regions of complex folding and shearing, i.e. Retrograde and the contact zones proximal to the undulating boundary of the Hores. Very few samples of Alma and regions of thin and discontinuous Cues remain after the isolation of unclassified samples. In comparison, SVM unclassified samples are spatially coincident with Cues in the west and Freyers, Parnell and Allendale classes in the south. The spatial distribution of unclassified samples is very different for RF than for SVM. RF associates high uncertainties to a large proportion of misclassified samples and contact zones.

Conversely, SVM identifies spatially contiguous regions of high uncertainty with particular classes and not necessarily to erroneous predictions or contact zones.

5.5. Discussion

In this section, we discuss the interpretations of results and then provide a survey of the implications and potential applications of the methods presented. Our results show that the identification and removal of incorrect predictions using uncertainty thresholds significantly increases RF prediction accuracy especially for small T_a sample sizes. In addition, uncertain predictions are, in many cases, spatially coincident with zones proximal to lithology contacts and areas of shearing and intense deformation. Further evidence for the relationship between RF unclassified samples and lithology contact zones is presented in Figure 5.9. This graph shows mean T_b uncertainty as a function of distance from mapped lithology boundaries. In general, there is an overall decrease in mean RF uncertainty with distance and with an increase in the number of T_a samples. Furthermore, these graphs show that as T_a sample size increases there is a shift from an approximately linear decrease to a logarithmic decrease in mean RF uncertainty with distance. This suggests that with an

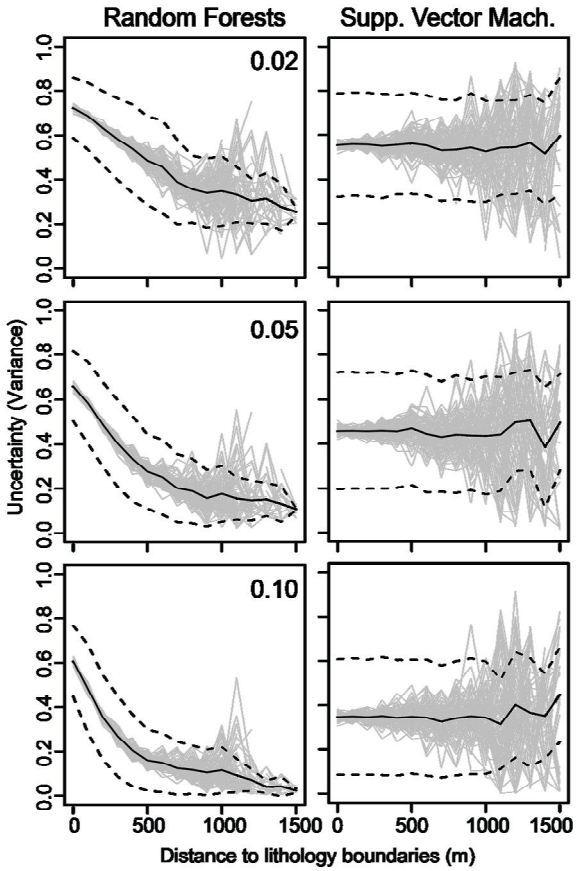

Figure 5.9 Comparison of RF and SVM T_b uncertainty as a function of distance in metres (100 m bins) to lithology contacts for 0.02, 0.05 and 0.10 T_a sample proportions. Thin grey lines = individual T_b. Black line = mean of T_b. Dashed line = one standard deviation from mean of T_b.

increase in the information content of the trained classification models, uncertain predictions are concentrated on or near contact zones. In contrast, Figure 5.9 shows high SVM uncertainties are not related to their distance from lithology boundaries. SVM is less likely to associate high uncertainty with misclassified samples which, in this example, are proximal to lithology contact zones. At distances greater than ~ 500 m the instability in mean uncertainty is due to a reduced number of samples available to construct these curves.

Identifying unclassified samples based on uncertainty thresholds can be used to substantially improve the overall accuracy of RF predictions. However, there is a trade-off between an improvement in RF accuracy and the elimination of a large number of samples. Our results show that SVM uncertainty, obtained from estimates of class membership probabilities using the PKPD method and calculated using the *Variance* metric, are not a good proxy for the identification of incorrect predictions. The most likely influences on SVM uncertainty are overfitting of its classification models to noisy T_a and high interclass similarities and intraclass variability. Take for instance, the Cues composite gneiss represents a wide variety of lithologies comprising psammopelitic to psammitic composite gneisses or metasediments, with intercalated bodies of basic gneiss (Willis *et al.* 1983). Moreover, confusion between Allendale, Cues, Freyers, Parnell and Sundown metasediments is likely to be a function of their relatively similar psammopelitic to psammitic metasediment compositions (Stevens *et al.* 1983; Willis *et al.* 1983). Investigations are required into other methods for calculating uncertainty from SVM class membership probabilities, such as the Best-versus-Second-Best (BvSB) approach outlined by Joshi *et al.* (2009). BvSB derives uncertainty based on the difference in class membership probabilities of the two classes with the highest values, therefore, it is not influenced by the probabilities of unimportance classes.

The source of RF uncertainty in our results likely to be a combination of two factors: the process of inferring discrete spatial features using data with inherently different *support*, i.e. geometries and/or resolution; and the presence of natural variability in geological phenomena. These factors result in noise or inconstancies within data (Hammer & Villmann 2007). Geophysical and geological data is unlikely to be neither completely precise nor entirely reproducible and thus contains some form of deterministic noise. In other words, the noise in the data we use to generate predictions contains information that we are either not interested in or cannot adequately model (Scales & Snieder 1998).

There are restrictions placed on predicting discrete spatial features that are a function of the geometry (size, shape and areal extent) or *support* of the individual measurements we have used to generate inferences (Goodchild & Quattrochi 1997). For example, there is an inherent minimum resolution of the samples used to represent a continuous field, i.e. a pixel. Given that the boundaries of contiguous pixels are unlikely to coincide with discrete and irregular contacts, we would expect that a pixel covering such transitions will contain information representing multiple classes. It is expected that output class membership probabilities should reflect the presence of mixed classes within a single sample, leading to an increase in uncertainty (Goodchild *et al.* 1992; Gotway & Young 2002).

Geological materials are rarely homogeneous in their composition. This natural variability leads to inter (between) class homogeneities and intra (within) class heterogeneities (Goodchild *et al.* 1992; Fisher 1999). For example, high RF uncertainty values were common for samples proximal to Retrograde Schist zones, the Hores and Alma gneisses. These classes share a common thin and discontinuous geometry. Furthermore, in the complexly folded central east of the sample area Retrograde schist is likely to be a mix of metamorphosed lithologies (Stevens *et al.* 1983). In contrast, Hores Gneiss exhibits an undulating and irregular contact with the underlying Freyers Metasediments, characteristic of highly metamorphosed regions (Willis *et al.* 1983). Classes that represent undifferentiated lithologies such as the Purnamoota Subgroup were often predicted with high uncertainty. Undifferentiated units are indicative of poor surface exposure due regolith cover resulting in the generalisation of several distinct lithologies.

There are many situations in applied geophysics and geosciences in general where identifying surface/sub-surface transition zones is more advantageous than the prediction of categorical values in isolation, especially where samples cannot be assigned to a specific class (Hammer & Villmann 2007). For example, this approach may be generalised to other geoscience disciplines such as: exploration and applied geophysics, hydrogeology and environmental geology; and geomorphology.

Regions of intense deformation, such as shear zones, can be associated with particular types of economic mineral deposits. For example, gold-bearing quartz-veins regularly occur in spatial proximity to shear zones (e.g., Kusky & Ramadan 2002; Jackson 2005; Holden *et al.* 2008; Metelka *et al.* 2011). As a precursor to the identification of mineral exploration targets, litho-structural maps are routinely constructed by integrating airborne

geophysics, geochemistry, satellite multi-spectral data and field observations. The RF supervised classification and uncertainty analysis methods presented here can provide these types of studies with a robust indication of the location of particular lithologies and associated measures of prediction confidence. This approach can be used to identifying regions containing interesting geological features, such as contact zones for prioritising subsequent field work in order to optimise survey objectives.

Draskovits & Laszlo (2005) and Buselli &Lu (2001) combined electromagnetic survey methods, hydraulic conductivity and geological structural data to image groundwater contamination. Qualitative interpretations of contamination zones were based on the combination of these complementary sources of data. Our method lends itself to the integration of multiple relevant variables, which as Draskovits & Laszlo (2005) conclude, provide more information than a single source of data. Uncertainty mapping could be used to expand these interpretations in order to highlight regions of rapid contamination migration and/or locations requiring further observations.

An example of a further likely application of this approach is in the identification of areas prone to landslides. Landslides are a significant threat to human life and infrastructure (Sabatakakis *et al.* 2012). Landslide hazard mapping is therefore vital for identifying regions susceptible to mass movement processes in order to mitigate their potential impacts. Approaches to landslide susceptibility assessment involve the combination of multiple sources of information including, lithology and geological structures, topography and associated derivatives, climate data and human infrastructure (e.g., Tangestani 2004; Ramli *et al.* 2010; Sabatakakis *et al.* 2012). The outputs of these assessments define zones of low to high risk based on predefined criteria. Our method could be used, for example, to identify structural lineaments and contacts zones for input into the hazard susceptibility mapping process. Alternatively, RF classification uncertainty could be used to flag ambiguous or mixed category features in the outputs of studies such as Stumpf & Kerle (2011), where RF was used in combination with object-based image analysis to classify landslide features.

Our demonstration infers the location of lithologies and identifies regions of uncertain classifications using the RF algorithm and geophysics/remote sensing data for a multiclass problem with a high-dimensional variable space. We have shown that the spatial distribution of RF uncertainty accurately identifies misclassified samples associated with

abrupt or irregular contact zones and the thin, discontinuous surface geometries of the most difficult to classify classes. Characterising the spatial dependencies between error, uncertainty and these potentially interesting geological features is beyond the scope of this work. However, the similarities between RF and *k*-Nearest Neighbours (kNN) classifiers (Hastie *et al.* 2009) provide an indication of why RF is able to exploit spatial dependencies in data and SVM is not. RF is essentially an adaptive form of *k*NN that assigns weights to T_a observations proximal to points requiring prediction using only the most informative variables.

5.6. Conclusions

Uncertainty estimates are an easily obtainable component of machine learning outputs. Our novel demonstration example compares Random Forests and Support Vector Machines for the task of inferring the spatial distribution of lithology in a complex, folded, high-grade metamorphic terrane from integrated airborne geophysics and satellite multispectral data. We show that the uncertainty associated with Random Forests categorical predictions identifies the majority of misclassified samples, especially when a limited amount of training information is available. Incorrect or ambiguous Random Forests predictions identified using uncertainty thresholds established from available test data were isolated, resulting in a significant improvement in overall prediction accuracies and individual class recall and precision rates of the remaining samples. In contrast, uncertainties estimated for Support Vector Machine predictions were not associated with incorrect classifications. Random Forests classification uncertainty, based on rasterised airborne and spaceborne data, is due to a combined influence of inferring discrete spatial features using data with inherently different support (i.e. geometry and resolution) and the presence of natural variability in geological phenomena. We show that Random Forests is, in this context, superior to Support Vector Machines with regard to exploitation of inherent dependencies and structures contained within spatially varying input data. Random Forests provides analysts with an easy to use and highly accurate method for inferring the spatial distribution of lithology and generating robust prediction uncertainties. The methodology presented here can be used to significantly improve Random Forests prediction accuracy and also to highlight regions containing significant geological features such as abrupt changes in lithologies related to transition or contact zones and regions of intense deformation and metamorphism.

5.7. Acknowledgements

Landsat ETM+ data was sourced from the United States Geological Survey and airborne geophysics data from Geoscience Australia. This research was conducted at the Australian Research Council (ARC) Centre of Excellence in Ore Deposit (CODES) under Project No. P3A3A. M. Cracknell was supported through a University of Tasmania Elite Research Ph.D. Scholarship. We thank two anonymous referees for their constructive comments that have strengthened the content and clarity conveyed within this manuscript.

CHAPTER 6 – MAPPING GEOLOGY AND VOLCANIC-HOSTED MASSIVE SULFIDE ALTERATION IN THE HELLYER–MT CHARTER REGION, TASMANIA, USING RANDOM FORESTS™ AND SELF-ORGANISING MAPS

Published in Australian Journal of Earth Sciences,
http://dx.doi.org/10.1080/08120099.2014.858081, 2014

6.0. Abstract

The Hellyer–Mt Charter region of western Tasmania includes three known and economically significant volcanic-hosted massive sulfide deposits. Thick vegetation and poor outcrop present a considerable challenge to ongoing detailed geological field mapping in this area. Numerous geophysical and soil geochemical datasets covering the Hellyer–Mt Charter region have been collected in recent years. These data provide a rich source of geological information that can assist in defining the spatial distribution of lithologies. The integration and analysis of many layers of data in order to derive meaningful geological interpretations is a non-trivial task, however, machine learning algorithms such as Random Forests™ and Self-Organising Maps offer geologists methods for identifying patterns in high-dimensional (many layered) data. In this study, we validate an interpreted geological map of the Hellyer–Mt Charter region by employing Random Forests to classify geophysical and geochemical data into 21 discrete lithological units. Our comparison of Random Forests supervised classification predictions to the interpreted geological map highlights the efficacy of this algorithm to map complex geological terranes. Furthermore, Random Forests identifies new geological details regarding the spatial distributions of key lithologies within the economically important Que–Hellyer Volcanics. We then infer distinct but spatially contiguous sub-classes within footwall and hangingwall, basalts and andesites of the Que–Hellyer Volcanics using Self-Organising Maps, an unsupervised clustering algorithm. Insight into compositional variability within volcanic units is gained by visualising the spatial distributions of sub-classes and associated statistical distributions

of key geochemical data. Compositional differences in volcanic units are interpreted to reflect contrasting primary composition and volcanic-hosted massive sulfide alteration styles. We conclude that combining supervised and unsupervised machine learning algorithms provides a widely applicable, robust means, of analysing complex and disparate data for machine-assisted geological mapping in challenging terranes.

Key words: Geological mapping, volcanic-hosted massive sulphide, machine learning, Random Forests, Self-Organising Maps, Tasmania.

6.1. Introduction

The Hellyer–Mt Charter region is located in west Tasmania (Figure 6.1) and hosts three economically significant volcanic-hosted massive sulfide (VHMS) Pb–Zn ore bodies: Que River, Hellyer and Fossey. VHMS mineralisation is restricted to a sequence of upper middle Cambrian marine calc-alkaline mafic to felsic volcanic facies known as the Que–Hellyer Volcanics (QHV, Corbett & Komyshan 1989; Waters & Wallace 1992; Gemmell & Fulton 2001). Although there is little Quaternary glacial cover in the region, detailed geological mapping is a challenging prospect due to dense temperate rainforest vegetation and thick soil profiles. The majority of field observations used to construct interpreted geologic maps were collected from road and track cuttings, in costeans and in the vicinity of high-voltage transmission lines (Corbett & Komyshan 1989). More recently, local and regional drill core data have been integrated with surface data as a means of improving mapping outcomes (e.g., McNeill *et al.* 1998; Gemmell & Fulton 2001; Tomes 2011). Due

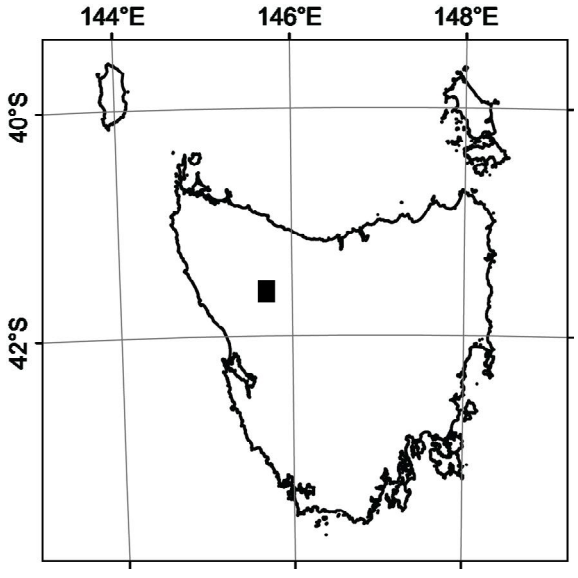

Figure 6.1 Map of Tasmania. Black rectangular region indicates the location of the Hellyer–Mt Charter region.

to the challenging conditions faced by mineral explorers in the Hellyer–Mt Charter region, methods that model the spatial distribution of surface/near-surface geology will be of significant value.

Since the early 1970s, numerous geophysical and geochemical datasets covering the Hellyer–Mt Charter region have been collected. These data have been used to assist geologists to characterise the complex geological setting of the economically significant QHV as means of targeting VHMS mineralisation (Corbett & Komyshan 1989; Richardson 1994; McNeill *et al.* 1998). However, the integration of many layers of data presents challenges with regard to their analysis and interpretation. MLAs provide practitioners with robust pattern recognition capabilities, especially where the dimensionality of input variables is high and where the relationships between data are poorly understood (Ripley 1996; Hastie *et al.* 2009; Kanevski *et al.* 2009). The rich but complex data available for the Hellyer–Mt Charter region provide an opportunity to employ MLAs to assess pre-existing geological interpretations and gain new insights into geological phenomena.

Supervised learning uses an inductive approach to identify patterns in data by constructing one or more predictive models that link input variables to target outcomes. In contrast, unsupervised learning utilises a data-driven approach to exploring and identifying natural groups (clusters) within data. These computational tools have the potential to provide users with information on interactions between data, so called data mining (Feyyad 1996; Ripley 1996; Witten & Frank 2005; Hastie *et al.* 2009; Kanevski *et al.* 2009; Marsland 2009). In this study, we combine training data (T_a) representing interpreted lithological units and input variables obtained from geophysical and geochemical surveys. We demonstrate the use of Random Forests (RF), in its capacity as a supervised classifier, to assess the validity of a pre-existing interpreted geological map of the Hellyer–Mt Charter region from these data. We then utilise Self-Organising Maps (SOM), an unsupervised clustering method, to identify if present, spatially contiguous and geologically meaningful sub-classes within individual volcanic units from key geophysical and geochemical variables. These computational tools are distinct but, when combined, facilitate our understanding of complex natural phenomena from large amounts of disparate data.

6.1.1. Geological setting

VHMS ore deposits constitute globally significant sources of Pb, Zn, Cu, Au and Ag. The world's major VHMS deposits are preserved in rocks displaying the characteristics of

extensional tectonic environments (Franklin *et al.* 2005; Galley *et al.* 2007). VHMS deposits have been grouped into five lithostratigraphic types based on the occurrence of different volcanic and sedimentary lithologies that formed simultaneously with VHMS mineralisation (Franklin *et al.* 2005): (1) bimodal-mafic; (2) mafic; (3) pelite-mafic; (4) bimodal-felsic; and (5) siliciclastic-felsic settings.

VHMS deposits form at or near the seafloor from interactions with metal-enriched hydrothermal fluids generated by seafloor convection systems (Large *et al.* 2001; Franklin *et al.* 2005; Galley *et al.* 2007). VHMS hydrothermal alteration zone mineralogy and composition are a direct result of the source of hydrothermal fluids, their temperature and pH and how these fluids interact with host units (Galley 1995). Changes in the controls on VHMS mineralisation are a function of the maturity of the hydrothermal convection system at the time of alteration (Franklin *et al.* 2005).

VHMS ore deposits in the Hellyer–Mt Charter region fall into the category of the bimodal-

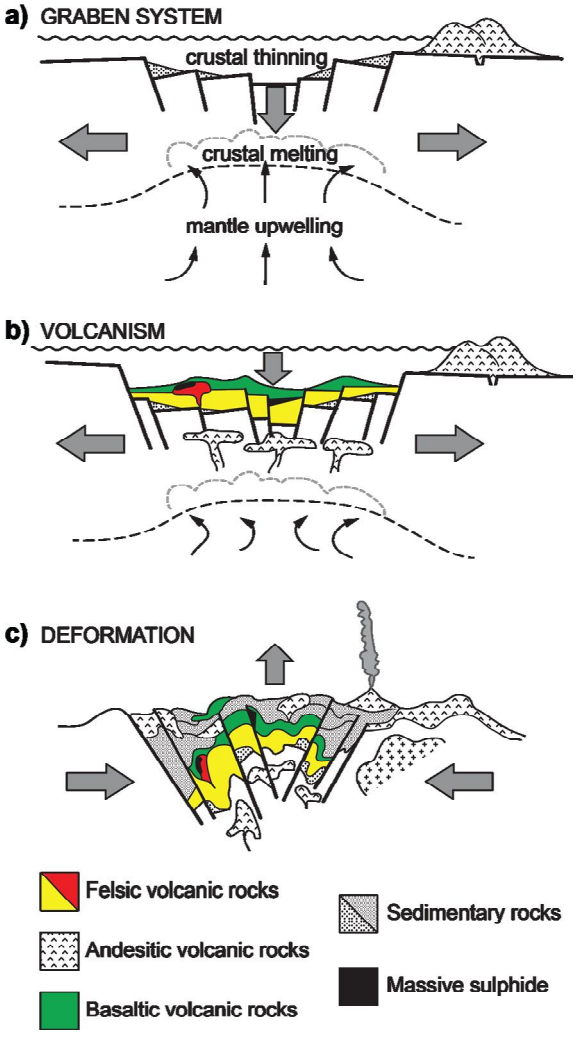

Figure 6.2 Typical regional tectonic and geological setting of bimodal-felsic VHMS ore deposits: a) graben development in a volcanic back-arc basin; b) volcanism and VHMS ore deposit formation; and c) compressive arc environment and deformation, from Allen et al. (2002).

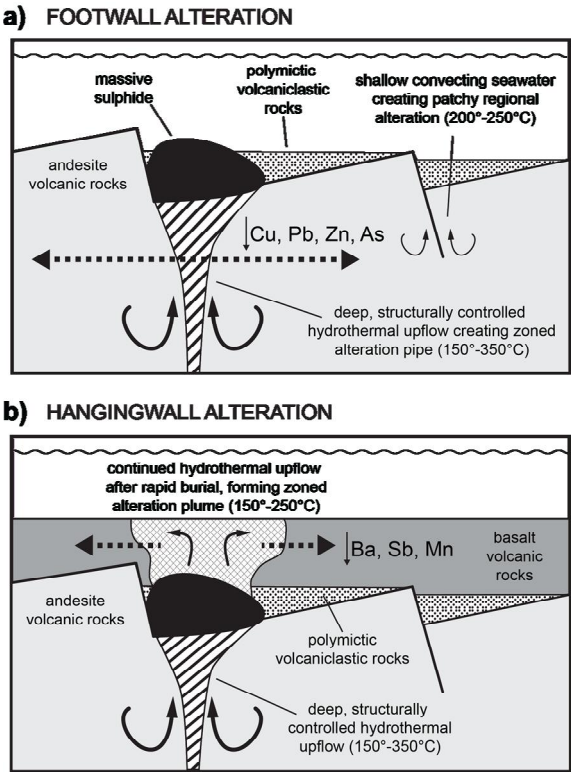

Figure 6.3 Diagram of VHMS a) footwall and b) hangingwall hydrothermal alteration zones in the Hellyer–Mt Charter region, modified from Gemmell & Fulton (2001). Dashed arrows indicate changes in abundance of mobile elements with distance from a) alteration pipe and b) alteration plume surrounding the Hellyer ore deposit.

felsic lithostratigraphic type (Franklin *et al.* 2005). Hellyer–Mt Charter VHMS mineralisation is interpreted to have formed in a back-arc tectonic setting as lenses of polymetallic massive sulfides (Figure 6.2, Allen *et al.* 2002). Distinctions have been made between Hellyer–Mt Charter VHMS footwall and hangingwall hydrothermal alteration systems based on differences in the spatial distributions of mobile elements (Figure 6.3, Gemmell & Fulton 2001).

The mechanisms that generate VHMS mineralisation coupled with post-mineralisation deformation result in complex geometries of VHMS host horizons and nearby units. Prospecting for VHMS mineralisation requires detailed, i.e. 1:20,000 scale, geological maps depicting the spatial distribution of lithologies and hydrothermal alteration zones (Franklin *et al.* 2005; Galley *et al.* 2007).

VHMS mineralisation in the Hellyer–Mt Charter region is hosted by a highly variable package of altered volcanic rocks within the QHV known as the Mixed Sequence (Table 6.1, Corbett & Komyshan 1989; Waters & Wallace 1992; McNeill *et al.* 1998; Gemmell & Fulton 2001). In addition to the three economic stratiform and stratabound VHMS ore deposits noted above, there is one sub-economic Au prospect at Mt Charter. The oldest rocks in the study area are feldspar-phyric volcanic rocks of the Central Volcanic Complex (CVC, Figure 6.4). In the study area, the Mt Charter Fault defines the contact between the

Table 6.1 Hellyer–Mt Charter lithological units and their stratigraphic relationships (Corbett & Komyshan 1989; Waters & Wallace 1992).

Age	Map Symbol	Group	Name	Description
Quaternary	**Qs**			alluvia/swamp deposits, clays sand and gravels
Tertiary	**Tb**			basalt lava
Early Ordovician	**COo**	Owen Group	Owen Conglomerate	pink siliciclastic conglomerate and sandstone
Cambrian	**R**			quartz–feldspar porphyry intrusions
	Rdi	Southwell Subgroup		rhyodacite coherent rock
	Rvc			rhyolitic volcaniclastic rock
	URS			Undifferentiated
	UR2			bedded to massive quartz-feldspar-phyric volcaniclastic rock
	UR1			greywacke-siltstone
	QRS		Que River Shale	black pyritic mudstone
	Dol			dolerite sills
	HB	Que$_1$ Hellyer Volcanics	Hellyer Basalt	amygdaloidal basalt sheet lava, pillow lava
	A		Mixed Sequence	andesite lava
	D			dacite flow banded lava and volcaniclastic rock
	Y			polymictic lapilli to breccia volcaniclastic rock
	HA			sericite-pyrite-quartz ± chlorite altered rock
	Afp			andesite feldspar-phyric lava and volcaniclastic rock
	LB		Lower Basalt	massive, pillowed and brecciated basalt and volcaniclastic rock
	ACG		Animal Creek Greywacke	micaceous greywacke and siltstone
	BHB		Black Harry Beds	shard-rich volcaniclastic siltstone with minor volcaniclastic sandstone
	CVC	Central Volcanic Complex	Mt Black Formation	feldspar-phyric volcaniclastic and coherent rocks

CVC and micaceous sandstone and siltstone of the Animal Creek Greywacke, the lower unit of which comprises volcaniclastic siltstone and sandstone, known as the Black Harry Beds (Corbett & Komyshan 1989). The Animal Creek Greywacke is conformably overlain by predominantly andesitic to basaltic volcanic units interbedded with minor sedimentary rocks known as the QHV.

The QHV contains four stratigraphic subdivisions. At the base of the QHV lies the Lower Basalt. This unit has highly variable thickness and outcrops in the south-eastern quadrant of the study area. Conformably overlying the Lower Basalt is a unit of feldspar-phyric andesite that underlies the majority of VHMS mineralisation. The Mixed Sequence comprises strongly altered rocks, polymictic volcaniclastic rocks, dacite, hangingwall andesite and minor basalt. Lateral facies and thickness variations are rapid within the

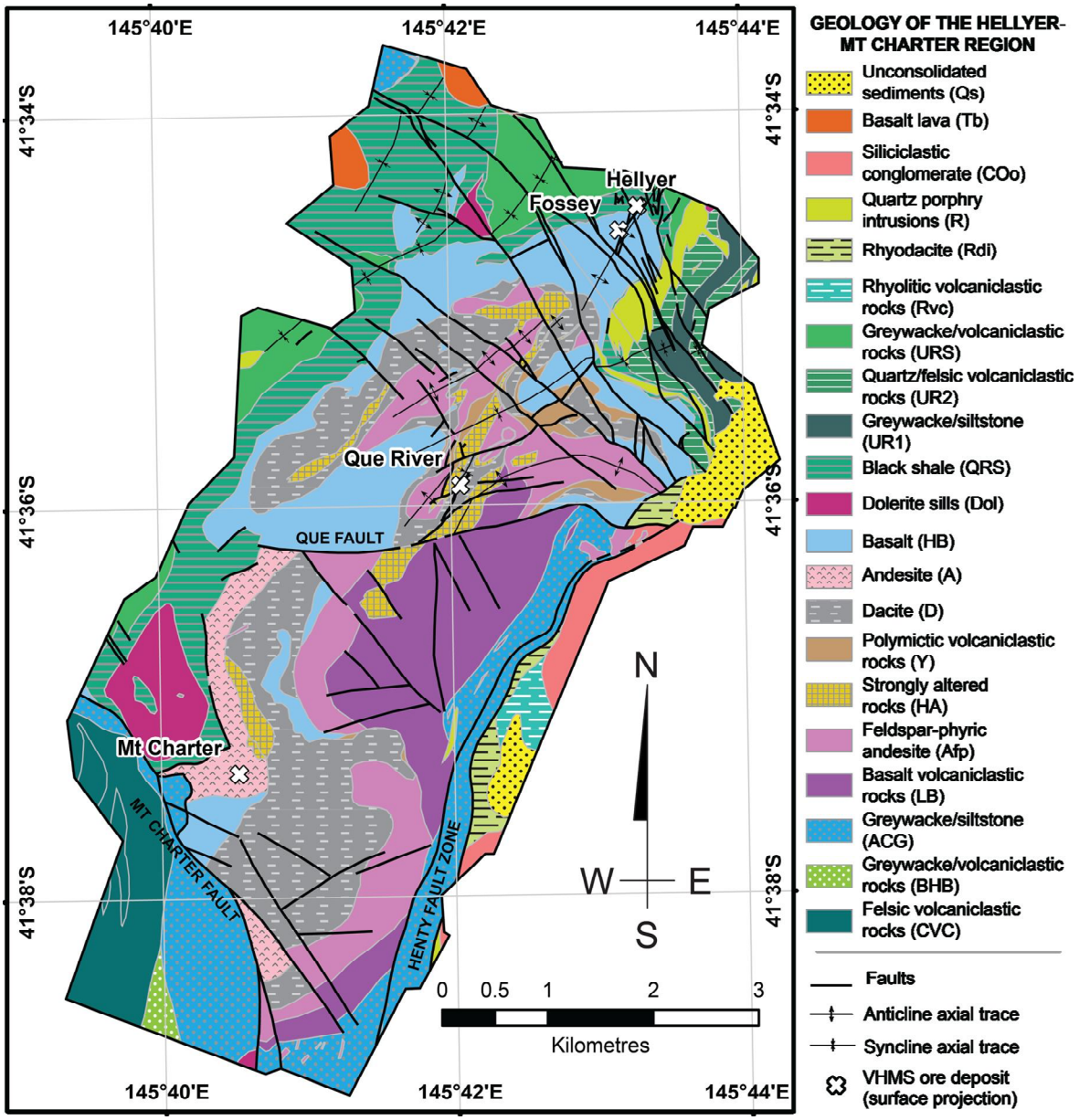

Figure 6.4 Interpreted geological map of Hellyer–Mt. Charter region, after Richardson (1994).

Mixed Sequence and polymictic volcaniclastic rocks are variably hydrothermally altered. The uppermost unit in the QHV is the Hellyer Basalt, a sequence of pillowed to massive basaltic to andesitic rocks (Waters & Wallace 1992).

The QHV are conformably overlain by the Que River Shale, a black pyritic mudstone that in some areas, appears to be intruded by the Hellyer Basalt. Rare dolerite sills, with similar composition to the Hellyer Basalt also intrude the Que River Shale in the south and north of the study area (Corbett & Komyshan 1989; Tomes 2011). Conformably overlying the Que River Shale is a sequence of volcaniclastic units, greywacke and quartz–feldspar porphyry sills of the Southwell Subgroup. In the study area, upper Cambrian–lower

Ordovician siliciclastic Owen Conglomerate unconformably overlies the Southwell Subgroup. Small areas of Cenozoic basalt lava occur in the north of the study region and obscure older rocks. Unconsolidated Quaternary alluvial and swamp deposits are found in the east of the study area at low altitudes (Corbett & Komyshan 1989).

Geological structures present within the Hellyer–Mt Charter region are a result of multiple overprinted tectonic environments and related events. A series of subsidence-related northeast-striking normal faults linked by northwest-striking transfer faults, probably of Cambrian age, are interpreted from magnetic and gravity surveys, rapid changes in facies and variations in unit thickness (McNeill *et al.* 1998). The Henty Fault Zone is a 100–300 m wide, north-trending zone of highly cleaved and lineated rocks, which was likely to have been active during the Cambrian, although the most recent movement along this fault post-dates Devonian folds (Berry 1989).

The Que Fault strikes east-northeast and offsets the Henty Fault Zone. The Mt Charter fault variably strikes southeast and forms the contact between older CVC, the younger Animal Creek Greywacke and the QHV. Changes in the thickness of volcanic and sedimentary units in the vicinity of the Mt Charter fault suggest it was active during the eruption and deposition of the QHV (Corbett & Komyshan 1989). During Devonian-aged metamorphism pre-existing faults were reactivated and folds trending north-northeast were generated. Fold plunges are in the order of 20–40 ° and fold wavelengths decrease from 2 km to approximately 500 m close to the Henty Fault Zone. Northwest-trending folds with wavelength of less than 400 m overprint north-northeast trending structures and generate dome and basin geometries (Corbett & Komyshan 1989).

6.1.2. Random Forests

RF, developed by Breiman (2001), is an ensemble MLA that utilises a majority vote cast by multiple random decision trees to generate class predictions. Unlike standard classifiers such as the Maximum Likelihood Classifier, which use a single classifier to generate predictions, ensemble classifiers train multiple classifiers and combine their results to generate predictions (Kuncheva 2004; Waske *et al.* 2012). Randomness is introduced into RF by randomly subsetting a predefined number of input variables (*mtry*) to split at each node of the Decision Trees (DT) and by bagging. Bagging (Breiman 1996) obtains T_a for DT by randomly sampling with replacement a number of samples equal to the number of instances in T_a. This equates to approximately two-thirds of the instances available for

training while the remaining samples are used for testing (Out-of-Bag samples). The Gini Index (Breiman *et al.* 1984) is employed by RF to calculate the information purity of child nodes as compared to that of their parent node. Split thresholds are determined from the maximum reduction in class heterogeneity, i.e. purity (Breiman 2001; Waske & Braun 2009). An estimate of the probability that a given sample is a member of one of *c* classes in T_a, i.e. class membership probabilities p_c, is calculated by dividing the frequency of votes for a given class by the number of DT in the forest (Liaw & Wiener 2002; Hastie *et al.* 2009). In this study, we quantify the degree to which the distribution of p_c is concentrated within a particular class using a modified version of the Kohavi & Wolpert (1996) variance metric. This metric generates consistent values of uncertainty between 0 and 1 regardless of the number of possible classes (Chapter 5, Cracknell & Reading 2013).

In Chapters 4 (Cracknell *et al.* 2014) and 5 (Cracknell & Reading 2013) of this thesis, RF was shown to perform well against other classification schemes with respect to predicting spatially distributed categorical features using multisource and high-dimensional remote sensing data (see also, Ham *et al.* 2005; Pal 2005; Waske *et al.* 2009; Ghimire *et al.* 2012). Of these studies, only three, Waske *et al.* (2009) and Chapters 4 (Cracknell *et al.* 2014) and 5 (Cracknell & Reading 2013) of this thesis, have used RF to classify lithology. Waske *et al.* (2009) took AVIRIS hyperspectral imagery of an unvegetated region covering the Hekla volcano in Iceland and compared RF with standard classifiers such as Maximum Likelihood Classifier (Mitchell 1994) and Spectral Angle Mapper (Kruse *et al.* 1993). Waske *et al.* (2009) showed that RF performed significantly better than standard classifiers. Chapters 4 (Cracknell *et al.* 2014) and 5 (Cracknell & Reading 2013) of this thesis, compared RF and other MLAs (e.g. Support Vector Machines, Vapnik 1998) to classify a sequence of sparsely vegetated Proterozoic high-grade metasedimentary and metavolcanic rocks of the Broken Hill region, Australia. In this study, airborne geophysical data and Landsat ETM+ multispectral imagery are used as input variables and T_a samples were distributed across the entire study region. Waske *et al.* (2009) and Chapters 4 (Cracknell *et al.* 2014) and 5 (Cracknell & Reading 2013) of this thesis found RF to be an attractive method for machine-assisted geological mapping. RF generated accurate results with a limited number of T_a samples, was stable in light of variations in classification model parameters and insensitive to irrelevant data. Furthermore, RF prediction uncertainties provided a good indication of ambiguous and/or erroneous classifications.

6.1.3. Self-Organising Maps

SOM (Kohonen 1982; Kohonen 2001) is a data mining tool that has the ability to highlight subtle relationships within high-dimensional and seemingly disparate datasets (Fraser & Dickson 2007). SOM has been shown to be useful for analysing, visualising and clustering multisource and high-dimensional geoscientific data (e.g., Penn 2005; Fraser & Dickson 2007; Bierlein *et al.* 2008; Bedini 2009; Carneiro *et al.* 2012). SOM employs the principles of vector quantisation and vector similarity measures, such as Euclidian distances, as a means of associating *d*-dimensional input samples to cells (nodes) mapped onto a 2D rectilinear grid (Kohonen 2001). The topology between nodes arranged on the 2D map indicates their relative proximities in *d*-dimensional variable space. SOM nodes summarise the variable characteristics displayed by discrete groups or clusters of input samples representing dominant patterns and structures present within the data (Bierlein *et al.* 2008).

SOM nodes are derived by assessing individual input samples using an iterative two-stage process. Firstly, an input sample is compared to randomly seeded nodes that fall within a predefined radius of assessment in variable space. The closest seed-node is deemed to be the winning seed-node. The properties of the winning seed-node are then adjusted by a percentage to approximate the characteristics of the nearest input sample to which it is deemed to be associated. Secondly, the characteristics of seed-nodes within a given radius of assessment to the winning seed-node are altered to more closely correspond to the properties of the input sample being scrutinised. These steps are repeated while reducing the radius of assessment in variable space and the percentage adjustment of seed-node properties. In this way, seed-nodes become nodes exhibiting variable characteristics representing clusters of input samples (Fraser & Dickson 2007; Bierlein *et al.* 2008).

6.2. Data and Methods

6.2.1. Source data

In this study, we used airborne geophysical, soil geochemical and Landsat ETM+ spectral radiance data (Table 6.2) covering an area of ~ 36 km^2. Appendix E provides detailed descriptions of data sources and data specific pre-processing methods employed in this study. Pre-processed data were transformed to a common coordinate system, resampled to 20 m grid pixel resolution (i.e. 90,429 samples) using bilinear interpolation, log transformed to approximate a normal distribution if required and scaled to zero mean and

unit variance. Figure 6.5 shows the spatial distribution of selected pre-processed datasets used in this study.

6.2.2. Data sampling

Supervised classifiers require representative T_a in order to adequately characterise the phenomena under investigation (Hastie *et al.* 2009). Using stratified random spatial sampling, we obtained class labels from a pre-existing interpreted geological map (Richardson 1994). This dataset contained 2100 samples, 100 samples for each of the 21 geological classes present in the study area and was used for RF classification model validation and training. We employed this method to obtain T_a because records of field observations used to construct the interpreted map were not available. Moreover, an equal number of T_a samples for individual classes reduce the potential for constructing a bias classifier. Bias classifiers arise where there is a large difference in the number of samples for each class present in T_a, leading to preferential prediction of those classes with large T_a

Table 6.2 Pre-processed input datasets (variables) used in this study.

Airborne Geophysics	Soil Geochemistry	Landsat ETM+
DEM [ab]	$\log_e$(Ag) [a]	Band 1
RTP [ab]	$\log_e$(As) [a]	Band 2
RTP 1VD [ab]	$\log_e$(Ba) [ab]	Band 3
K [a]	$\log_e$(Cu) [ab]	Band 4 [a]
Th [a]	$\log_e$(Cr) [ab]	Band 5
U	$\log_e$(Ni) [ab]	Band 6 [a]
$\log_e$(K/Th) [a]	$\log_e$(Pb) [ab]	Band 7
U/Th [a]	Ti [ab]	Band 3/1
U^2/Th	Zn [a]	Band 3/2 [a]
$\log_{10}$(AEM 880)	$\log_e$(Zr) [ab]	Band 3/5 [a]
$\log_{10}$(AEM 980) [a]	Ti/Zr	Band 3/7
$\log_{10}$(AEM 6k)		Band 5/1
$\log_{10}$(AEM 7k)		Band 5/2 [a]
$\log_{10}$(AEM 34k)		Band 5/4
		Band 5/7 [a]
		Band 5/4 × Band 3/4

[a] Indicates non-redundant variables identified from correlation analysis. [b] Indicates relevant variables identified via ranked-variable selection method. Airborne geophysics data abbreviations: DEM = Digital Elevation Model; AEM = Airborne Electro-Magnetics; RTP = Reduced-To-Pole Total Magnetic Intensity; and RTP 1VD = RTP 1st Vertical Derivative.

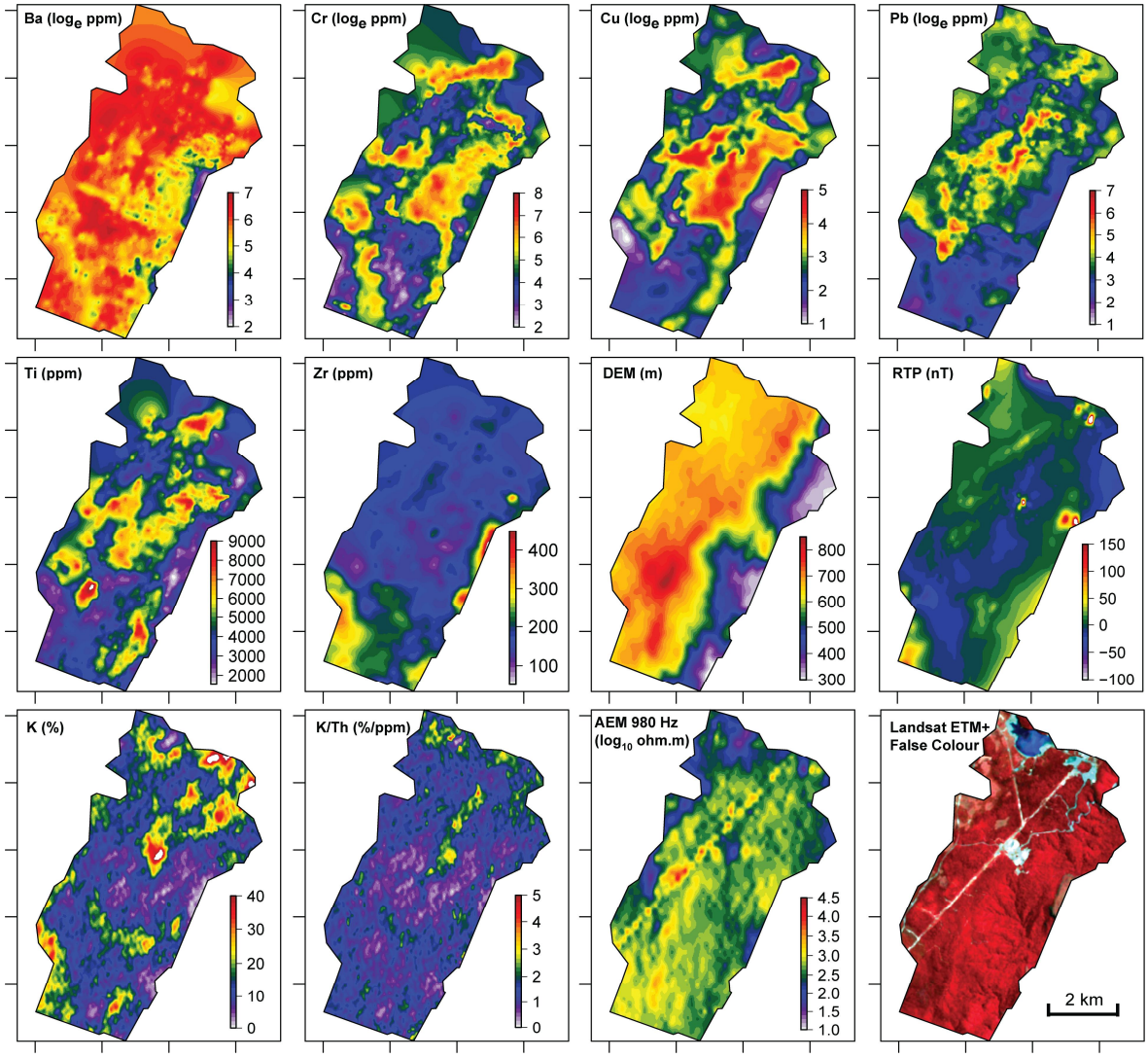

Figure 6.5 Examples of pre-processed (non-standardised) soil geochemical, airborne geophysical and Landsat ETM+ data used in this study. The Landsat ETM+ false colour image was generated by combining bands 4 (Red), 3 (Green) and 2 (Blue). Note the dense rainforest vegetation (red coloured) obscuring surficial geological materials.

proportions (Provost & Fawcett 1997; Japkowicz & Stephen 2002). Figure 6.6a shows the spatial distribution of T_a samples across the study area.

Test data (T_b), independent of T_a, were used to evaluate the performance of the trained RF classification model with respect to the interpreted geological map. T_b contained an equivalent number of samples to T_a (2100 or ~ 2.3 % of the total number of samples). In contrast to T_a, T_b were randomly sampled across the entire study area. Thus, T_b contained class sample proportions approximately equal in proportion to the area that a given class covers (Figure 6.6b).

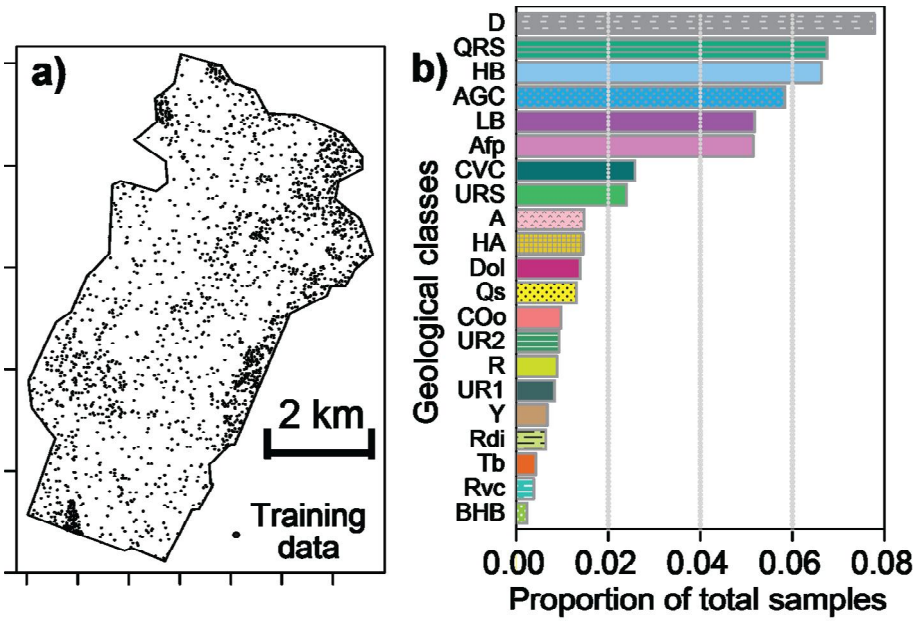

Figure 6.6 a) Location of the 2100 T_a samples used to optimise RF classification model parameters, select relevant variables and train RF classification models and b) individual class proportions of the total number of samples within the Hellyer–Mt Charter region, which approximates the proportions of classes in T_b. Class label abbreviations are given in Table 6.1.

6.2.3. Training Random Forests and variable selection

RF™, trademark of Leo Breiman and Adele Cutler, was implemented using the R package *randomForest* (Liaw & Wiener 2002). There are two RF parameters that require optimisation for a given supervised classification problem (Breiman 2001; Liaw & Wiener 2002), the number of possible variables to split at each node of a tree (*mtry*) and the number of decision trees to construct (**T**). We maintained **T** at the default value of 500 as the resulting error, obtained using Out-of-Bag samples, was observed to reach a stable minimum. The optimal *mtry* parameter for this classification task was selected based on the maximum mean accuracy obtained from the results of 10-fold cross-validation resampled 10 times.

MLAs are expected to discriminate between classes present within T_a of sufficient quality (Hastie *et al.* 2009). The inclusion of redundant or noisy data leads to overfitting and reduces the ability of classifiers to accurately discriminate classes. Therefore, methods that identify and select relevant variables will improve classifier performance (Yu & Liu 2004). The selection of relevant variables for this study was carried out in two stages. Firstly, highly correlated variables (see Table 6.2) with mean Pearson's correlation coefficients > 0.8 associated with a large proportion of other data were deemed to contain redundant or

duplicated information and eliminated. Duplicate information encourages supervised MLAs to place emphasis (if all variables are equally weighted) on redundant information to train classification models (Wettschereck *et al.* 1997; Guyon 2008). Secondly, a minimum number of relevant inputs were identified (see Table 6.2) using a ranked-variable selection method (Kuhn *et al.* 2012).

RF measures of variable importance were used to rank variables. RF quantifies variable importance, a measure of the impact of a variable on classification accuracy, based on the average decrease in Gini Index at decision tree nodes split using the variable in question (Liaw & Wiener 2002; Cutler *et al.* 2007; Waske *et al.* 2012). The ranked-variable selection method iteratively estimates the cross-validation accuracy for the top ranked *d* variables. The selection of a minimum number of relevant variables was based on the minimum number of top ranked variables with mean cross-validation accuracy within a user defined threshold of 0.015 from the maximum mean cross-validation accuracy across all combinations of variables. Figure 6.7a depicts the cross-validation accuracies obtained by the ranked-variable selection process. Figure 6.7b indicates the ranked importance of the 11 selected variables as measured by the mean decrease in child node heterogeneity (Gini Index) when splitting on that variable. These variables were used to train the final RF classification model employed to predict classes representing lithological units.

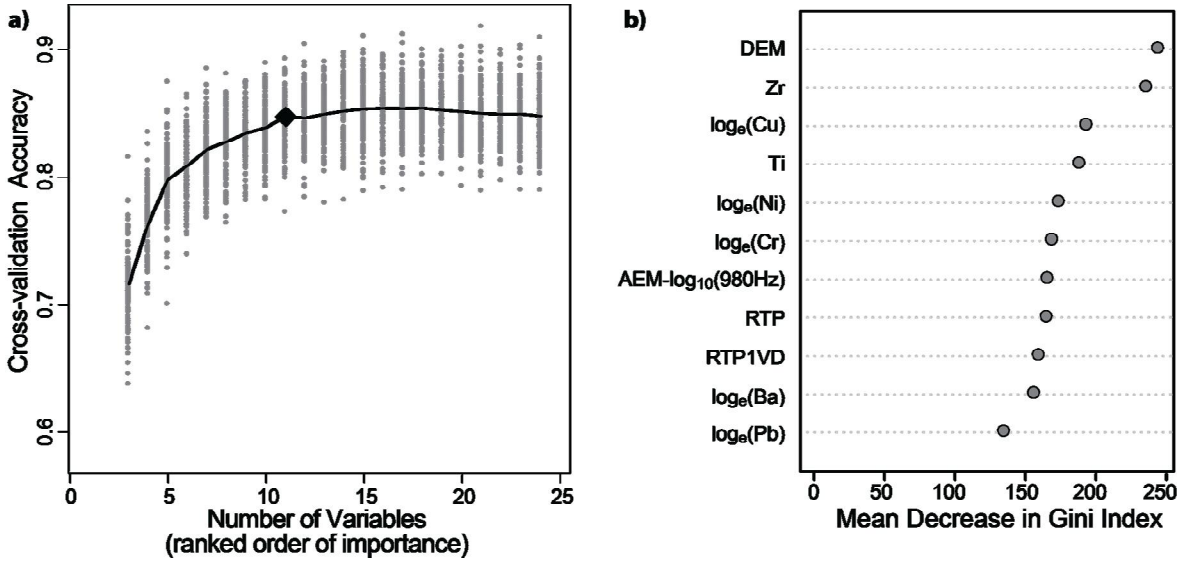

Figure 6.7 a) RF variable selection cross-validation accuracy. Grey points represent estimated 10 x 10-fold cross-validation accuracy. Black line plots mean cross-validation accuracy for each possible number of variables. Black diamond indicates the selected number of variables. b) Final list of selected variables in ranked order of relative importance based on mean decrease in RF Gini Index.

*Table 6.3 Comparison of parameter selection results of the different stages of variable selection, * denotes selected mtry values based on maximum mean accuracy for the 10 resampled 10-fold cross-validation iterations.*

No variable selection (41 variables)				**Non-redundant variables (24 variables)**				**Variable selection (11 variables)**			
mtry		Accuracy	1 × st. dev.	mtry		Accuracy	1 × st. dev.	mtry		Accuracy	1 × st. dev.
2		0.835	0.020	2		0.847	0.019	2	*	0.840	0.021
6	*	0.843	0.019	4	*	0.849	0.020	3		0.839	0.021
10		0.841	0.021	6		0.846	0.019	4		0.837	0.021
14		0.839	0.021	9		0.845	0.022	5		0.835	0.022
18		0.838	0.021	11		0.843	0.022	6		0.832	0.022
23		0.836	0.021	13		0.842	0.022	7		0.829	0.022
37		0.834	0.021	16		0.841	0.023	8		0.829	0.021
31		0.833	0.022	18		0.84	0.023	9		0.827	0.022
35		0.831	0.026	20		0.837	0.023	10		0.826	0.021
40		0.829	0.023	23		0.837	0.023				

Table 6.3 compares RF mean and standard deviation cross-validation accuracy estimates before and after the two-stages of variable selection. There is < 0.01 difference between the resulting maximum mean cross-validation accuracies, which is well within the ranges of their standard deviations. Although using a minimum number of relevant variables does not constitute the highest mean cross-validation accuracy, it does represent the minimum number of variables before mean accuracy decreases rapidly. The use of a minimum set of meaningful variables speeds up processing and assists in the comprehension of MLA decision structures (Kotsiantis 2007).

The most important variables used in this study are geophysical and geochemical data. The Digital Elevation Model (DEM) was the most important variable, indicating that lithological units are positioned at specific locations (and thus elevations) in the regional landscape. Immobile elements, such as Ti, Zr and Cr are useful for identifying primary igneous rocks (Hallberg 1984) and have been used with success to discriminate among QHV units (Corbett & Komyshan 1989; Crawford *et al.* 1992). Of the mobile elements, elevated Cu and Pb levels are associated with regions close to footwall VHMS alteration, whereas, Ba is more commonly found in elevated levels within hangingwall alteration zones (Gemmell & Fulton 2001). Airborne Electro-Magnetic (AEM) data have been used with relative success to target VHMS ore bodies and Total Magnetic Intensity (TMI) data

are useful for identifying regions of magnetite destruction associated with hydrothermal alteration (Richardson 1994).

Variables excluded from the final classification model constitute noise or data that does not, in this case, assist in the classification of lithological units. For example, Gamma Ray Spectrometry (GRS) data have been shown to be of little use in discriminating among geological units of the QHV (Richardson 1994). Furthermore, dense vegetation obscures geological spectral characteristics and thus, Landsat ETM+ data will only be of use in regions with limited vegetation cover (Grebby *et al.* 2011).

RF predictions were reclassified using a majority convolution filter that assigns the most abundant class within a 3 × 3 neighbourhood to the centre pixel. This type of post-classification filter is common in spatial classification problems and has been shown to improve prediction accuracy by removing the bulk of classifications representing high-frequency noise (Ghimire *et al.* 2010).

6.2.4. Implementing Self-Organising Maps

SOM was implemented using the R package *kohonen* (Wehrens & Buydens 2007). Samples (pixels) classified as one of the four basalt and andesite QHV units by RF were analysed. SOM parameters (*kohonen* package defaults) used to train nodes were: number of iterations 100; initial search radius two-thirds the dimensions of variable space; linear radius decrease; and percentage adjustment of seed-node properties from 0.05 to 0.01. SOM nodes were randomly seeded and projected onto a 3 × 3 hexagonal map. A small number of nodes were deliberately chosen because we wanted to group input samples into a minimal number of spatially contiguous and geologically relevant clusters.

Hierarchical dendrograms (Ripley 1996) and unified-distance matrices (U-Matrix, Ultsch & Vetter 1994) were used to visualise similarities between SOM nodes and identify an optimal number of merged SOM nodes (Vesanto & Alhoniemi 2000). Hierarchical dendrograms provide a 1D representation of the similarities between SOM nodes and were used to identify geologically relevant sub-classes representing hard clusters of samples associated with a given volcanic unit. Hierarchical dendrograms utilise an agglomerative method to combine nodes at each level of the dendrogram by iteratively merging the two closest (most similar) nodes, based on Euclidian distances in variable space, until the user defined number of node clusters is obtained (Ripley 1996; Witten & Frank 2005).

We plot the spatial distributions of SOM sub-classes in conjunction with the statistical distributions of their individual variable characteristics. These plots offer an opportunity to assess if spatially coherent patterns and thus meaningful results (Fraser & Dickson 2007) were generated during SOM analyses and also provide a means of interpreting sub-class geological significance.

6.3. Results

6.3.1. Geological classification using Random Forests

Our RF predictions of classes representing lithological units provide new geological detail for the Hellyer–Mt Charter region. Table 6.4 presents a confusion matrix assessing RF predictions with T_b. Overall T_b accuracy, i.e. the number of correct predictions divided by the total number of predictions, obtained for this confusion matrix is 0.784 ± 0.018 based on exact 95 % Confidence Intervals. Figure 6.8 (a, b) shows a comparison of the interpreted geological map and predicted lithological classes for all samples from the Hellyer–Mt Charter region. Abrupt transitions between lithologies at boundaries mapped as regional faults, such as the Que Fault (locality A in Figure 6.8b) and Mt Charter Fault, are well defined by RF. In addition, RF depicts accurately major lithologies within the QHV. For example, the predicted distribution of dacite matches its interpreted extent closely enough to identify fold geometries (locality B in Figure 6.8b). RF predictions imply that the interpreted Hellyer Basalt outcrop east of Mt Charter (locality C in Figure 6.8b) is likely to be a mixture of basalt and andesite intrusions located on the contact between feldspar-phyric andesite and dacite units.

Observed inconsistencies between the pre-existing interpreted geological map and RF predictions (Figure 6.8c) indicate locations where new geological insights may be implied. Examples of such informative mismatches are concentrated on or near contacts between Mixed Sequence lithologies. In particular, Mixed Sequence units located between the Hellyer and Que River VHMS deposits show a large degree of disagreement. Discrepancies are coincident with a region of pervasive hydrothermal alteration and where small faults, perpendicular to the Henty Fault Zone, generate spatially irregular lithologies. In addition, Que River Shale is commonly misclassified as Hellyer Basalt, dolerite and the undifferentiated shale-rich unit of the Southwell Sub-Group. This misclassification is due to strong basalt chemical signatures within the Que River Shale (Sinclair 1994), its spatial

proximity to dolerite and the high shale content of the undifferentiated Southwell Sub-Group.

The majority of differences between predictions and the reference map occur between stratigraphically and spatially proximal units within the Mixed Sequence, however, strongly altered rocks and polymictic volcaniclastic rocks are not misclassified as each other. Table 6.5 shows that strongly altered rocks and polymictic volcaniclastic rocks both exhibit precisions < 0.5. Low precision implies that the majority of predictions for these classes differ from the interpreted map. Nonetheless, these units display recall rates > 0.8, which indicates the majority of T_b samples for these classes are classified correctly. In contrast, the majority of other units within the QHV display higher precision than recall. Significantly lower precisions obtained for predictions of altered units suggest that VHMS

Table 6.4 T_b confusion matrix for RF predictions. Cells in the confusion matrix represent counts of class predictions for reference samples of a particular class. Counts of correct classifications for a given class are indicated along the diagonal. Note grey cells represent counts for correct predictions. Box indicates classes contained within the Que–Hellyer Volcanics. Classes are in chronological order and decrease in age from top to bottom and left to right. Class label abbreviations are given in Table 6.1.

		Prediction																			
		CVC	**BHB**	**ACG**	**LB**	**Afp**	**HA**	**Y**	**D**	**A**	**HB**	**Dol**	**QRS**	**UR1**	**UR2**	**URS**	**Rvc**	**Rdi**	**R**	**COo**	**Tb**	**Qs**
Reference	**CVC**	92	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	BHB	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	ACG	5	1	201	3	5	0	0	1	2	0	0	1	0	0	1	1	8	0	5	0	1
	LB	0	0	6	168	17	6	8	2	0	1	1	0	0	0	0	0	0	0	0	0	0
	Afp	0	0	5	6	128	12	19	13	2	3	0	1	0	0	0	0	0	0	0	0	0
	HA	0	0	0	0	3	43	0	4	1	0	0	0	0	0	0	0	0	0	0	0	0
	Y	0	0	0	0	2	0	20	2	0	0	0	0	0	0	0	0	0	0	0	0	0
	D	0	0	1	0	15	33	18	231	9	11	0	1	1	0	0	0	0	1	0	0	0
	A	0	0	3	0	0	0	0	0	58	0	0	0	0	0	0	0	0	0	0	0	0
	HB	0	0	0	2	6	2	17	10	12	187	0	9	9	1	0	0	2	5	0	0	0
	Dol	0	0	1	0	2	0	0	0	0	0	52	0	0	0	0	0	0	0	0	0	0
	QRS	0	0	1	0	0	1	0	5	10	28	17	181	1	1	23	0	0	8	0	10	0
	UR1	0	0	0	0	0	0	0	0	0	0	0	0	24	5	0	0	0	4	0	0	1
	UR2	0	0	0	0	0	0	0	0	0	0	0	0	2	24	0	0	0	0	0	0	2
	URS	0	0	0	0	0	0	0	0	0	0	1	2	0	0	75	0	0	1	0	0	0
	Rvc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0
	Rdi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	1	0	0
	R	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	27	0	0	0
	COo	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0
	Tb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0
	Qs	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	55

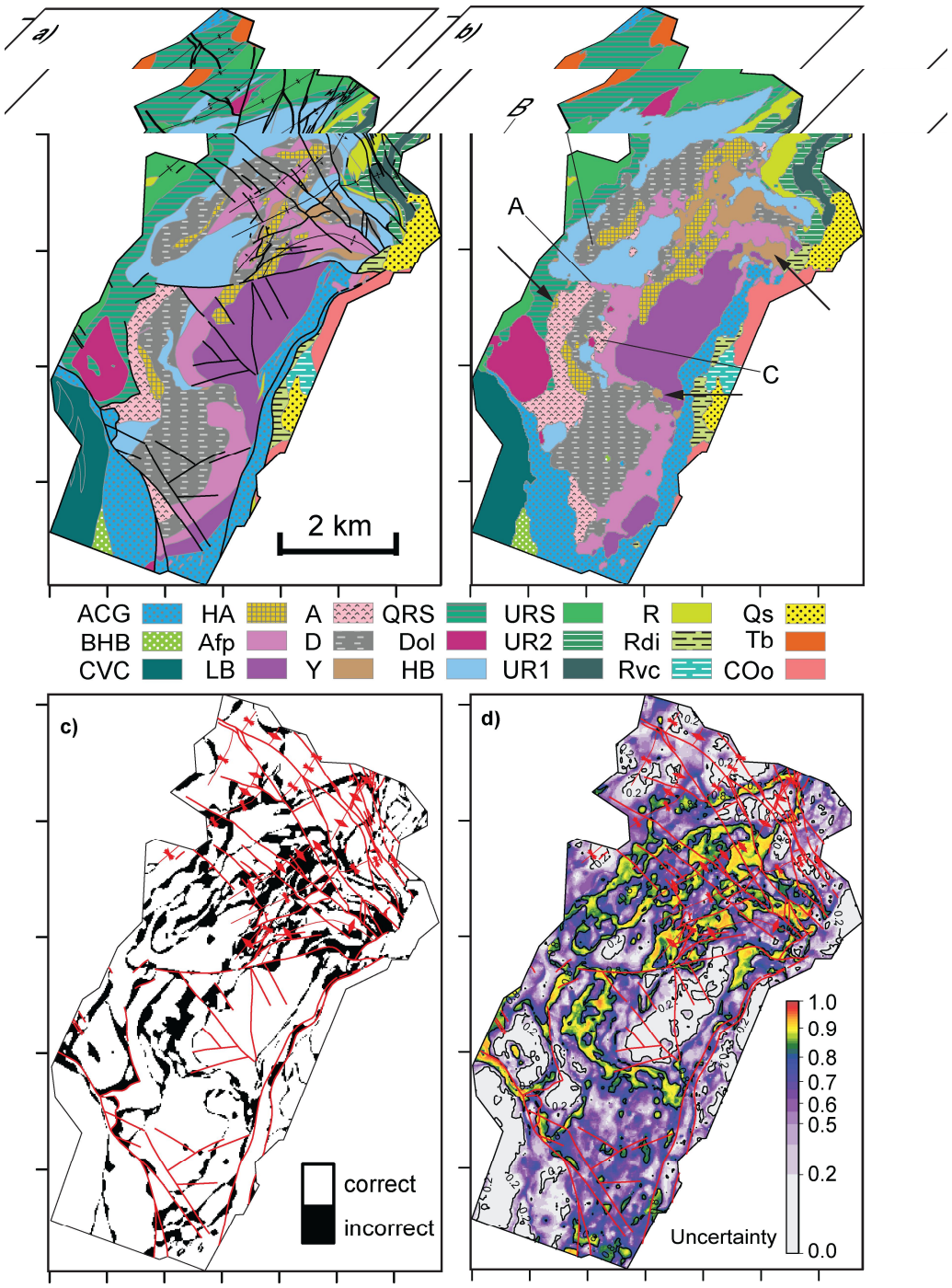

Figure 6.8 Comparison of: a) interpreted geological map, after Richardson (1994); b) lithology predictions generated by RF, arrows indicate previously unmapped bodies of strongly altered rock and polymictic volcaniclastic rocks that reflect potential sub-surface VHMS mineralisation and upper case letters refer to descriptions in the main text; c) spatial distribution of inconsistencies between the pre-existing interpreted geological map and RF predictions; and d) spatial distribution of RF prediction uncertainty, 0.2 and 0.8 contours indicated and natural logarithm colour scale applied. Note the close correspondence between regions of misclassified samples and uncertainty values greater than ~ 0.8. Red symbols in c) and d) represent faults and fold axial traces as defined in Figure 6.2. Class label abbreviations are identical to those presented in Table 6.1.

alteration is more widely distributed than was originally mapped.

The spatial distribution of RF prediction uncertainty is provided in Figure 6.8d. Warm colours indicate values > 0.8, which represents high uncertainty and implies that multiple classes are candidates for prediction. A comparison of Figure 6.8 (c, d) shows that high uncertainty is coincident with the majority of mismatched samples. These ambiguous classifications are concentrated at, or near, transition zones between classes and imply that additional data are required to constrain accurately the spatial distribution of lithologies in these areas. For example, the mapped extent of Que River Shale and altered and faulted QHV units between the Hellyer and Que River deposits are predicted with high uncertainties.

Table 6.5 Comparison of RF T_b recall and precision rates calculated from the confusion matrix in Table 6.4. Recall represents an estimate of the probability that a reference sample is correctly classified, whereas precision is an estimate of the probability that a prediction is correct (Congalton & Green 1998). Note class abbreviations are given in Table 6.1

	Class	Recall	Precision
	Qs	0.932	0.932
	Tb	1.000	0.583
	COo	0.962	0.807
	R	0.871	0.587
	Rdi	0.955	0.618
	Rvc	1.000	0.941
	URS	0.949	0.750
	UR2	0.857	0.727
	UR1	0.706	0.632
	QRS	0.633	0.928
	Dol	0.946	0.732
QHV	**HB**	0.714	0.813
	A	0.951	0.617
	D	0.720	0.862
	Y	0.833	0.244
	HA	0.843	0.443
	Afp	0.677	0.719
	LB	0.804	0.939
	ACG	0.855	0.905
	BHB	1.000	0.800
	CVC	0.979	0.949

6.3.2. Discrimination of geological sub-classes using Self-Organising Maps

Using SOM we were able to identify new lithological sub-divisions, represented by spatially contiguous regions with distinct geochemical composition, within individual volcanic units predicted by RF. We derived SOM clusters for basalt-dominated and andesite-dominated QHV units using nine key geophysical and geochemical variables. These variables represent the same datasets used to generate RF predictions excluding the DEM and Reduced-To-Pole 1st Vertical Derivative. These variables were omitted from SOM analyses as our motivation was to identify geochemical and/or geophysical variability within volcanic units.

Figure 6.9 shows four arbitrarily labelled sub-classes associated with Hellyer Basalt based on dendrogram analysis of SOM nodes. Both the dendrogram (Figure 6.9a) and U-Matrix (Figure 6.9b) imply that Hellyer Basalt sub-classes HB_1 and HB_3 are similar to each other, as are HB_2 and HB_4. The spatial distribution of Hellyer Basalt sub-classes (Figure 6.9c) indicates HB_4 is a narrow and discontinuous body along the lower margin of the Hellyer Basalt. HB_2 occurs in close proximity and stratigraphically above HB_4 north of the Que Fault. HB_1 is positioned as two regions within the hinge zone of a double plunging fold in the northwest of the study area. HB_3 occurs as a discrete body in the hinge zone of a syncline south of Mt Charter (not mapped in Figure 6.4). The largest differences between HB_1 and HB_3 and HB_2 and HB_4 are apparent in Ti, Cu, Cr and Pb variables (Figure 6.9d). HB_1 and HB_3 display relatively higher levels of Ti, Cu and Pb compared to HB_2 and HB_4. HB_1 contains the highest Cu and HB_3 the highest Pb of the four sub-classes, whereas HB_4 contains the lowest Cr. Similarities in immobile elements (Zr, Ti, Cr and Ni) are represented by HB_2 and HB_4. As implied by the relative abundance of mobile elements (Cu, Pb and Ba), HB_1 and HB_3 indicate regions likely to be affected by VHMS alteration.

Figure 6.10 presents SOM derived sub-classes for Lower Basalt, feldspar-phyric andesite and andesite units. These figures contain maps representing the spatial distribution of sub-classes derived from dendrogram merging of SOM nodes and frequency density estimates for the variables that contribute most to the dissimilarities between sub-classes. We find that Lower Basalt (Figure 6.10a) sub-classes LB_1 and LB_2 occur adjacent to each other immediately south of the Que Fault. LB_3 and LB_4 are mapped as two neighbouring but distinct bodies close to the Henty Fault Zone in the southeast of the study region. LB_1 and

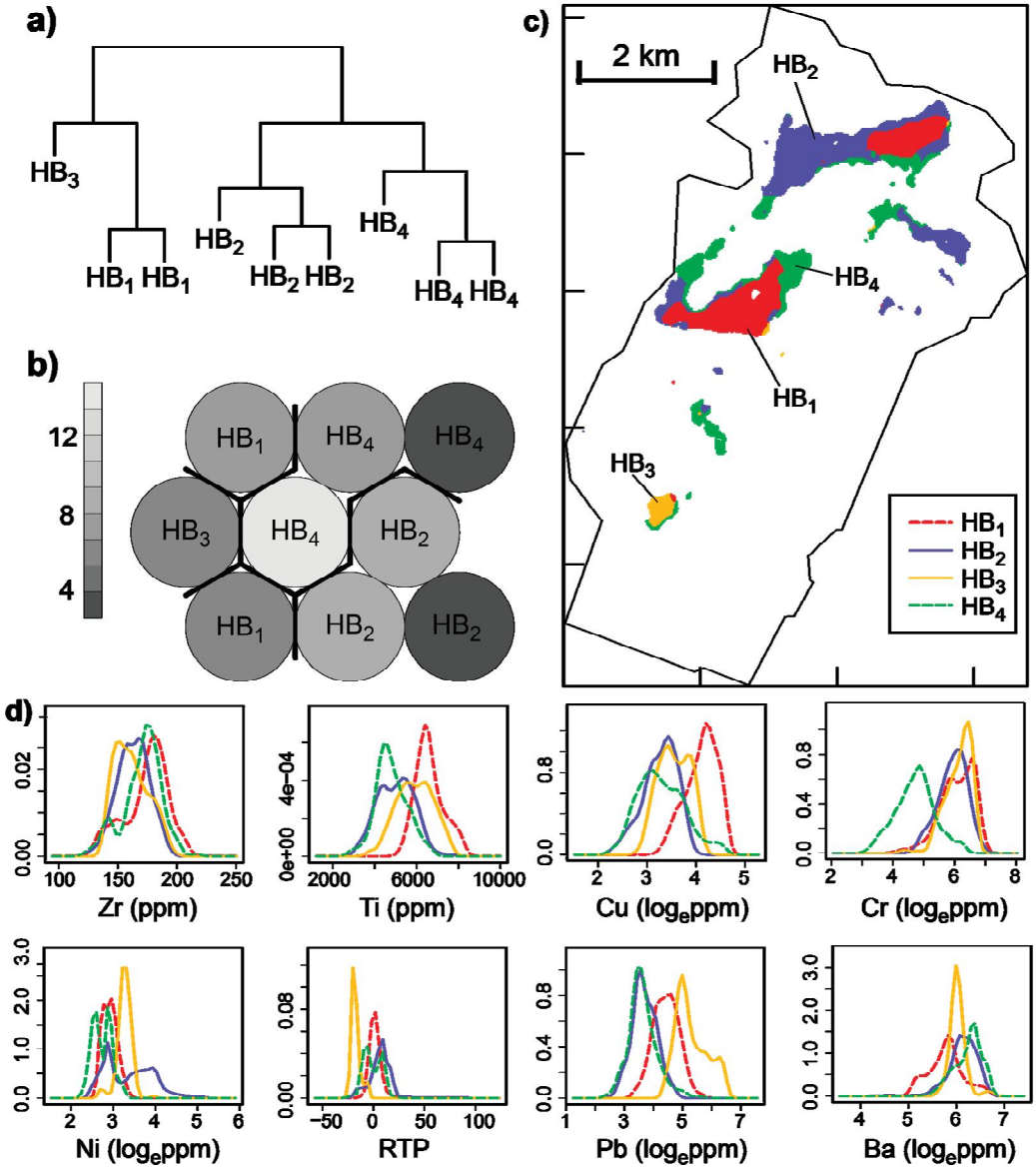

Figure 6.9 Results of SOM unsupervised clustering for the Hellyer Basalt divided into four sub-classes: a) hierarchical dendrogram sub-classes representing clustered SOM nodes; b) U-matrix representing distance to neighbours of nine SOM nodes, lighter colours indicate nodes with larger Euclidian distances in variable space to neighbouring nodes, the resulting sub-class membership of SOM nodes are overlayed; c) spatial distribution of Hellyer Basalt sub-classes; and d) frequency density estimations for eight key input variables used to derive sub-classes. The y-axes of variable frequency plots represent frequency densities.

LB_2 have similar Zr, Ti, Cr and Ni values, whereas LB_2 contains elevated Pb levels compared to the other sub-classes.

The feldspar-phyric andesite and andesite units of the QHV were both divided into three sub-classes. Figure 6.10b shows feldspar-phyric andesite sub-class Afp_1 is positioned as a small region adjacent to an elongate body of Afp_3 close to the southern extent of the Henty Fault Zone. Afp_2 is mapped as two separate bodies, one to the south of the Que Fault and the other located northeast of Que River. North of the Que Fault, Afp_1 and Afp_3 occur as separate bodies east and west of Afp_2. Afp_1 and Afp_3 exhibit similar (high) Zr and (low) Pb distributions, whereas major differences are observed in Ti, Cu and Ba abundance. Afp_2 is

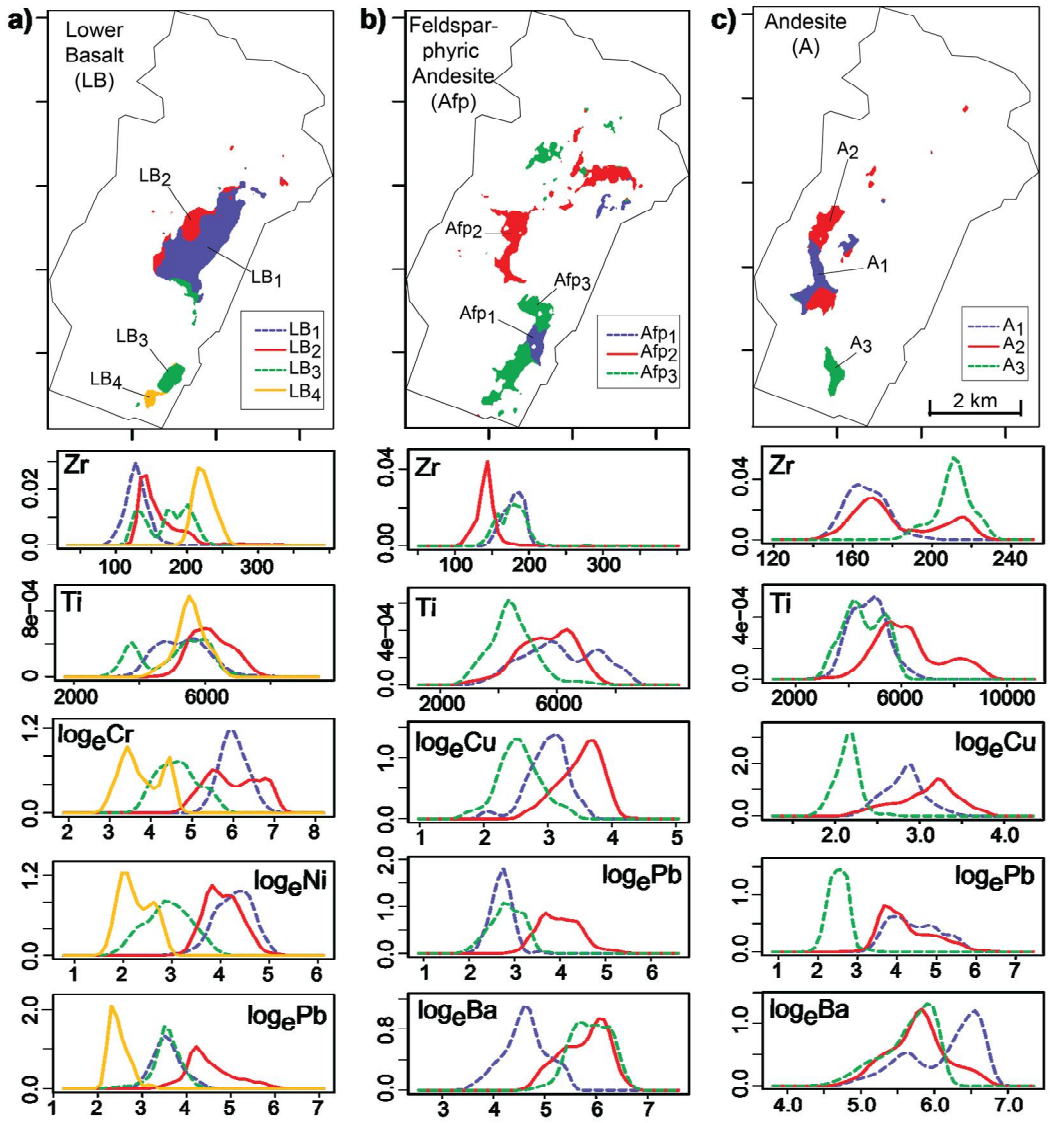

Figure 6.10 SOM sub-classes identified (using the same methodology as described for Hellyer Basalt) in a) Lower Basalt, b) feldspar-phyric andesite and c) hangingwall andesite units of the QHV. The y-axes of variable frequency plots represent frequency densities.

characterised by high values of Pb and Cu.

Figure 6.10c shows the hangingwall andesite sub-class A_1 is located in the centre of an elongate body of andesite straddling the Mt Charter deposit. A_1 is in immediate contact to strongly altered rocks (Figure 6.8b) and is characterised by elevated levels of Cu and Pb (both approximately equal to A_2) and Ba. A_2 is located as two separate bodies surrounding A_1 and exhibits relatively high Cu values. The relatively low concentration of mobile elements observed in A_3 suggests that it is not close to sources of VHMS hydrothermal alteration.

6.4. Discussion

Rocks exhibiting strong hydrothermal alteration are key exploration targets for VHMS deposits in the Hellyer–Mt Charter area. Altered rocks at the surface are likely to be connected to sub-surface zones that are prospective for VHMS mineralisation (Franklin *et al.* 2005). Recall and precision rates derived from the analysis of RF predictions indicate the probability that strongly altered and polymictic volcaniclastic units are correctly classified is higher than the probability that predictions of these units are incorrect. Based on these findings, we can infer the presence of several previously unmapped regions of prospective VHMS alteration (indicated with arrows in Figure 6.8b). Mineralisation targets are located south of the Que Fault on the western margin of the body of andesite, as small bodies east of Mt Charter and as an extension to the polymictic volcaniclastic rocks east of Que River adjacent to the Henty Fault Zone.

The ratio between the "immobile" elements Zr and Ti provides a means of identifying compositional differences in igneous rocks (Hallberg 1984; Crawford *et al.* 1992). Observed differences in Zr/Ti within Lower Basalt sub-classes imply a shift from basaltic to andesitic dominated composition with distance from the Que Fault. Furthermore, geochemical similarities and spatial proximities of LB_1, LB_2 and Afp_2 indicate compositional relationships among these sub-classes. The Lower Basalt and feldspar-phyric andesite are stratigraphically the oldest units of the QHV and were already in place during VHMS mineralisation (Corbett & Komyshan 1989). LB_2 and Afp_2 display the greatest potential for intense footwall hydrothermal alteration as they represent sub-classes with relatively high values of Pb. Furthermore, Afp_2 is found in close proximity to the majority of predictions of strongly altered rock north of the Que Fault.

Hangingwall andesite sub-class A_1 is characterised by high Ba, indicating a region of hangingwall alteration (Gemmell & Fulton 2001) close to predictions of strongly altered rock. This region of hangingwall alteration is implied to extend eastwards toward the small body of A_1. In order to generate hangingwall alteration plumes in volcanic units overlying VHMS mineralisation, hydrothermal convection systems must continue for some time after the (intrusive or extrusive) emplacement of these units (Gemmell & Fulton 2001).

Distinctions between SOM node clusters representing primary compositional and alteration-affected sub-classes of the Hellyer Basalt suggest a lack of hangingwall VHMS alteration plumes in this unit at the surface north of the Que Fault. In contrast, the elevated Cu abundances of HB_1 and HB_3 are more likely to be linked with VHMS footwall alteration zones (Gemmell & Fulton 2001). Given that the Hellyer Basalt had been emplaced after episodes of footwall alteration (Waters & Wallace 1992), another mechanism must be responsible for the elevated levels of Cu within these sub-classes. The observed spatial distributions of HB_1 sub-classes are coincident with the axial traces of an anticline and syncline and directly above the Hellyer and Fossey VHMS deposits. Similarly, HB_3 is located close to the axial trace of a syncline (not mapped) south of Mt Charter and displays high Pb concentrations. In these sub-classes, Cu and Pb derived from footwall alteration zones were probably remobilised during Devonian deformation and redistributed via faults, fractures and veins. This interpretation provides an explanation for the observed concentration of footwall alteration mineralisation in areas within and adjacent to regions of high strain, i.e. fold hinges, that acted as foci for the transport of metal-rich fluids.

This study demonstrates a combination of supervised and unsupervised learning approaches that is widely applicable to geological data mining problems. There are an increasing number of publically available, high-resolution remote sensing geophysical and geochemical datasets covering the majority of the Australian continent, including GRS, TMI and geochemical abundance layers derived from ASTER satellite data (Cudahy *et al.* 2012). These data provide rich sources of information that can, with expert knowledge and understanding of the methods described herein, be integrated to validate interpreted geological maps and identify subtle but meaningful compositional differences within lithologies. The outputs of such studies will provide geoscientists with additional geologically relevant information for use in practical applications such as tectonic

reconstructions (e.g., Dohm *et al.* 2007; Gibson *et al.* 2008) and mineralisation prospectivity mapping (e.g., Binbin *et al.* 2011; Carranza 2011; Merdith *et al.* 2013).

6.5. Conclusions

The computational approach to integrated geochemical and geophysical data analysis described in this study has identified areas of significant interest for geological and volcanic-hosted massive sulfide alteration mapping in the Hellyer–Mt Charter region, Tasmania. The new information we have extracted may be summarised as follows:

1. The spatial extent of lithological units, previously mapped using sparse surface outcrop observations, has been successfully recovered using Random Forests, leading to new details regarding key volcanic-hosted massive sulfide mineralisation host lithologies.

2. Insights into geological processes are further gained through examination of small inconsistencies between the original geological map and predictions derived from remotely sensed data. Points of note include (a) the expansion of Mixed Sequence units representing zones of intense hydrothermal alteration and (b) the close relationship between Que River Shale and interleaved Hellyer Basalt.

3. Geological meaningful sub-classes within individual volcanic units were identified using Self-Organising Maps. Sub-classes within Hellyer and Lower Basalts reveal compositional variations overprinted by hydrothermal alteration related to volcanic-hosted massive sulfide systems and the remobilisation of geochemical elements during subsequent metamorphism. Sub-classes derived for andesitic units delineate contrasting zones of footwall and hangingwall volcanic-hosted massive sulfide alteration.

Machine learning algorithms provide geologists with opportunities to add new detail to existing geological maps, validate geological interpretations and identify subtle but informative variations within key lithologies. The computational methods demonstrated in this study are widely applicable to geological investigations where detailed field mapping is hindered by inaccessible terrain and poor outcrop.

6.6. Acknowledgements

We thank Jocelyn McPhie for her constructive comments on a draft version of this manuscript. Airborne geophysics data were sourced from Mineral Resources Tasmania, Landsat ETM+ data were sourced from the United States Geological Survey and soil geochemistry data were obtained from an Aberfoyle Resources Limited legacy dataset. This research was conducted at the Australian Research Council Centre of Excellence in Ore Deposits (CODES) under Project No. P3A3A. M. Cracknell was supported through a University of Tasmania Elite Research Ph.D. Scholarship. A. McNeill publishes with the permission of the Director of Mines, Mineral Resources Tasmania. Random Forests™ is a trademark of Leo Breiman and Adele Cutler. We thank Frank Bierlein and an anonymous reviewer for comments that improved the clarity of the manuscript and the language used to describe the methods and interpretations.

CHAPTER 7 – SPATIAL-CONTEXTUAL MACHINE LEARNING SUPERVISED CLASSIFIERS: LITHOSTRATIGRAPHY CLASSIFICATION EXAMPLE

To be submitted to IEEE Transactions on Geoscience and Remote Sensing.

7.0. Abstract

Machine learning algorithms naturally treat each input sample statistically independent and identically distributed. Hence, the spatial context of geospatial data is not addressed by standard machine learning algorithm classifiers. The incorporation of spatial-contextual information into supervised classification models is an emerging field of research in the geosciences. This chapter describes a series of experiments designed to implement and evaluate novel new and pre-existing methods that provide implicit spatial context to machine learning algorithms. Spatial-contextual methods are assessed with respect to a challenging supervised lithostratigraphy classification problem. Based on existing research, methods for providing spatial-contextual information can divided into three categories: (1) the use of focal operators to calculate 1st and 2nd order spatial statistics, "texture variables", as a means of transforming the original input variables; (2) the induction of multiple localised supervised classifiers using the principles of *k*-Nearest Neighbours to subset proximal training samples in variable space; and (3) post-regularisation of classifications using image segmentation (via unsupervised clustering algorithms) and majority focal operators. Exhaustive combinations of these spatial-contextual methods are evaluated with respect to their overall accuracy and interpretability in conjunction with a comparison between Random Forests™ and Support Vector Machines supervised classifiers. In order to facilitate these experiments, code was written in the R statistical programming language using a combination of modifications to existing functions and the development of new functions.

7.1. Introduction

Machine learning algorithms (MLAs), such as Random Forests (RF), Support Vector Machines (SVM) and Self-Organising Maps (SOM) offer practitioners methods for the integration of large disparate datasets for either supervised classification or unsupervised clustering applications (Ripley 1996; Witten & Frank 2005; Hastie *et al.* 2009; Marsland 2009). MLAs have seen increasing use in applications for multiclass lithology classification using remotely sensed geophysical data (e.g., Bedini 2009; Kovacevic *et al.* 2009; Bedini 2012; Carneiro *et al.* 2012; Yu *et al.* 2012). These studies and those documented in Chapters 4 (Cracknell & Reading 2014), 5 (Cracknell & Reading 2013) and 6 (Cracknell *et al.* 2014) of this thesis, have demonstrated the efficacy of MLAs, in particular RF, to predict classes representing lithological units in remote or inaccessible terranes from a limited number of training data (T_a). The development of methods that address spatial dependency and non-stationarity within the MLA workflow, henceforth called spatial-contextual classifiers, constitute an emerging field of research in geological remote sensing applications (Shaheen *et al.* 2011).

Standard practice for the classification of remote sensing images using MLAs treats spatially distributed samples (pixels) as statistically independent and identically distributed (Burl *et al.* 1998; Demšar *et al.* 2013). In these situations, MLA classifier training and testing is carried out on data using a single global pixel-based classification model (Zortea *et al.* 2007; Tarabalka *et al.* 2009). As a result, spatially distributed MLA predictions often contain a large amount of high-frequency noise characterised by “speckled” classifications. This noise has a detrimental effect on MLA performance and inhibits the interpretation of MLA outputs in the spatial domain. This classification noise is a function of the high interclass similarities and intraclass variability that is characteristic of lithological units (Ghimire *et al.* 2010; Grebby *et al.* 2011) and the inclusion of irrelevant or erroneous geophysical data (Ricchetti 2000; Link & Blundell 2003).

Global pixel-based MLA classifiers do not attempt to exploit the spatial heterogeneity and spatial dependencies commonly encountered in spatial processes (Atkinson & Tate 2000; Gahegan 2000; Fotheringham *et al.* 2002; Lloyd 2011; Demšar *et al.* 2013). Spatial heterogeneity refers to a spatial process that exhibits contrasting, non-stationary local statistical properties across a given domain or region. In these situations, different parameters are required to adequately characterise non-stationary processes or models at

various locations (Anselin 1995; Fotheringham *et al.* 2002; Lloyd 2011; Demšar *et al.* 2013). Spatial dependency, first described by Tobler (1970), refers to the characteristic of spatially distributed variables to display autocorrelation, i.e. measurements that are close together are more likely to be similar than those farther apart (Fotheringham 2009; Getis 2010; Lloyd 2011; Demšar *et al.* 2013).

Most spatial processes are scale dependent. Different scales of measurement or analysis will conclude that the process under investigation is homogeneous at a given scale and that it is heterogeneous at another (Atkinson & Tate 2000; Lloyd 2011). The effects that scale have on spatial data present challenges to their meaningful integration and analysis (Gotway & Young 2002). Adequately representing spatial process statistical characteristics may require changing the scale of the domain in which it is analysed (Franklin *et al.* 1996). In its simplest form, local spatial analysis utilises subsets of global/regional data, defined by local neighbourhoods surrounding the point of interest. These local neighbourhoods are used to generate multiple statistics or parameters across the spatial domain under investigation (Fotheringham *et al.* 2002; Getis 2010; Lloyd 2011).

Spatial-contextual information is either explicit or implicit depending on whether absolute or relative spatial context is used to train classifiers and obtain predictions. Kovacevic *et al.* (2009) and the experiments conducted in Chapter 4 (Cracknell & Reading 2014) included spatial coordinates as input variables for the supervised classification of lithology. These studies found that explicit spatial information reduced the degree to which classifications were affected by high-frequency noise. Nonetheless, explicit spatial information is most useful in situations where T_a are distributed across the entire region under investigation (Gahegan 2000; Cracknell & Reading 2014). Alternatively, implicit measures of spatial context provide MLAs with spatial information that is not bound explicitly to a geographical reference frame.

Recently, a range of methods have been developed that attempt to exploit implicit spatial-context for supervised classification of land cover and geomorphic features from remotely sensed data (e.g., Li & Narayanan 2004; Stepinski & Bue 2006; Zortea *et al.* 2007; Blanzieri & Melgani 2008; Tarabalka *et al.* 2009; Ghimire *et al.* 2010; Ghosh *et al.* 2010; Murray *et al.* 2010; Li *et al.* 2012; Segata *et al.* 2012). These methods can be divided into three general approaches that implicitly provide local spatial context to MLAs at different stages of the inference process: (1) those that attempt to characterise spatial context via

data pre-processing using low-pass filters and/or textural derivatives; (2) methods that exploit local spatial structures and generate multiple localised classifiers from subsets of training data; and (3) post-processing (post-regularisation) methods that reclassify predictions to represent spatially homogeneous regions.

7.1.1. Pre-processing methods

Pre-processing input variables prior to MLA training addresses the problem of characterising spatial-context in two ways (Gahegan 2000; Tarabalka *et al.* 2009; Lloyd 2011; Li *et al.* 2012): (1) use of focal operators to derive new values representing local similarity (or dissimilarity), i.e. texture, within local neighbourhoods; and (2) segmenting the original spatial domain into contiguous regions representing similar spectral characteristics. These two approaches can be viewed as either local (focal operators) or localised (segmentation) methods for obtaining spatial-contextual information directly from input variables. Local methods generate a single model derived from the characteristics within a local neighbourhood assuming non-stationary spatial processes. Localised methods are characterised by a group of models representing non-stationary spatial processes in discrete regions. Localised methods are commonly used to reduce computational cost by reducing the number of models required to adequately represent the spatial processes under investigation (Hengl 2009).

7.1.1.1. Focal operators

The most common method for analysing local spatial characteristics uses focal operators (moving windows, convolution filters). The output of focal operators is a function of the values of neighbouring pixels and can represent either 1st or 2nd order spatial statistics (Franklin *et al.* 1996; Shankar 2009; Murray *et al.* 2010; Lloyd 2011). 1st order spatial statistics represent properties calculated from neighbouring pixel intensity histograms, e.g. mean, variance, etc. Whereas, 2nd order spatial statistics offer methods for assessing the relationship between pairs of neighbouring pixel values such as texture derivatives, e.g. the outputs of Haralick *et al.'s* (1973) Grey-Level Concurrence Matrices (GLCM, Shankar 2009; Murray *et al.* 2010; Lloyd 2011).

Focal operator 1st order statistics have been used to provide some notion of spatial context to MLAs for the supervised classification of land cover, vegetation and lithology from remote sensing data (e.g., Franklin *et al.* 1996; Ricchetti 2000; Li & Narayanan 2004; Lepistö *et al.* 2006; Zortea *et al.* 2007; Shankar 2009; Ghimire *et al.* 2010; Murray *et al.*

2010; Grebby *et al.* 2011). Ghimire *et al.* (2010) incorporated a measure of spatial autocorrelation, derived from the Getis statistic (G*, Getis & Ord 1992; Ord & Getis 1995), as an input variable for RF classifier training and prediction of vegetation classes. The G* is a standardised focal operator that incorporates mean global values with low-pass (mean) values derived from local neighbourhoods (Ghimire *et al.* 2010). Different G* neighbourhood sizes were compared with a RF global pixel-based classifier post-regularised using majority focal operators (see Section 7.1.3) with different dimensions. Larger neighbourhoods were shown to generate the most accurate classifications (Ghimire *et al.* 2010).

2nd order spatial statistics, e.g. texture derivatives, can be thought of as a measure of the variability or regularity within a local region (Gonzalez & Woods 2008). There have been many studies specifically related to the classification of lithology using texture data. Zortea *et al.* (2007) randomly selected GLCM textures from different neighbourhood sizes to train multiple SVM classifiers. In this example, the contributions of trained SVM models were weighted to generate predictions. Ricchetti (2000) and Grebby *et al.* (2011) integrated multispectral reflectance imagery with topographic derivatives to classify lithology. The inclusion of topographic derivatives increased classification accuracies when compared to those resulting solely from classifiers trained on multispectral imagery. Li and Narayanan (2004) derived textures using Gabor wavelet filters. Lepistö *et al.* (2006) convolved multi-resolution Gaussian ring filters with images of drill hole core to derive statistical summaries of the mean and standard deviation of the transform coefficients and combined these into an ensemble *k*-Nearest Neighbours (kNN) classifier (Ricchetti 2000; Grebby *et al.* 2011). Shankar (2009) employed 2nd order spatial statistics to obtain texture derivatives from airborne magnetics data to classify lithologies.

7.1.1.2. Image segmentation

Gahegan (2000) acknowledged that an efficient and reasonable way to utilise spatial information without including coordinates was to employ segmentation (called tessellation). Segmentation divides the region under investigation into zones displaying homogeneous variable characteristics (Tarabalka *et al.* 2009). The simplest method for segmenting an image is to divide the region up into quadrants of equal area (Gahegan 2000; Lloyd 2011). Alternatively, multiple segmented regions of arbitrary size and shape can be derived using unsupervised clustering. Unsupervised clustering algorithms identify groups or clusters of samples in variable space reflecting zones of homogeneous spectral

characteristics (Stepinski & Bue 2006; Tarabalka *et al.* 2009; Ghosh *et al.* 2010). This is akin to the methods used in object-based classification, although object-based methods often require the provision of complex user defined rules to segment image space (Stow 2010; Stumpf & Kerle 2011; Duro *et al.* 2012). Unsupervised clustering algorithms, such as SOM (Kohonen 2001), have been used to highlight subtle relationships in and extract signatures related to the spatial distribution of rock types jointly derived from multiple geophysical datasets (Bedini 2009; Bedini 2012; Carneiro *et al.* 2012). In a similar study, Paasche & Eberle (2009) employed the fuzzy *c*-means clustering algorithm (Bezdek 1981) to integrate and segment images of geophysical data for mineral exploration applications.

7.1.2. Training data selection

Bottou & Vapnik (1992) proposed a simple local approach to machine learning supervised classification that utilised the *k*-nearest T_a samples to a T_b sample requiring prediction and then trains a linear learning algorithm. This method does not change the architecture of the learning algorithm just the training procedure. Nonetheless, significant gains in classifier accuracy were obtained, when compared to a standard kNN classifier, for a hand-written digit recognition problem. More recently, Blanzieri & Melgani (2006), Zhang *et al.* (2006) and Blanzieri & Melgani (2008) proposed a kNN-SVM hybrid classifier that used principles similar to those developed by Bottou & Vapnik (1992) described above. The kNN-SVM hybrid classifier identifies the *k*-nearest T_a samples in variable space to the sample requiring prediction. The kNN-SVM classifier predicts a class label for sample in question using a SVM classifier trained on the *k*-nearest T_a samples instead of using a majority vote based on *k* class labels (Blanzieri & Melgani 2006; Blanzieri & Melgani 2008; Segata *et al.* 2012).

The kNN-SVM method outlined above is computationally expensive because the number of trained classifiers is equal to the number of T_b samples. A reduction in processing time can be obtained using variations proposed by Segata & Blanzieri (2009) and Segata & Blanzieri (2010), whereby, for each T_a sample the *k*-nearest T_a samples in variable space are identified and used to train a localised classification model. Predictions for unknown samples are obtained using the local classification model centred on the closest T_a sample in variable space to the sample in question.

7.1.3. Post-processing methods

Spatial-contextual post-processing methods involve the reclassification of pixels based on spatially proximal classifications, so called post-regularisation (PR, Tarabalka *et al.* 2009). PR represents that reclassification of pixels such that relatively homogeneous regions of a single class result. PR has been used often in remote sensing lithology classification applications (e.g., Ricchetti 2000; Link & Blundell 2003; Toumani 2003; Grebby *et al.* 2011). This is because lithological units usually cover spatially contiguous regions larger than the scale of a single pixel. Therefore, it is unlikely that a classified sample will differ markedly to its immediate neighbours unless there is a transition from one lithological unit to another. Rapid and isolated classifications representing different classes are usually a response to high-frequency noise present within input variables (Ricchetti 2000; Link & Blundell 2003). The most commonly used PR method for remote sensing classification is the majority focal operators (Tarabalka *et al.* 2009; Ghimire *et al.* 2010; Li *et al.* 2012). Majority focal operators simply assign the most abundant class within a neighbourhood to the centre pixel.

7.1.4. Combination methods

Combined spatial-contextual methods incorporate two or more elements of the approaches described above. For example, utilising texture derivatives in conjunction with segmentation (Stepinski & Bue 2006; Ghosh *et al.* 2010) or coupling textural derivatives with PR majority focal operators (Ricchetti 2000; Grebby *et al.* 2011). An alternative approach proposed by Tarabalka *et al.* (2009), combines the outputs of global pixel-based classifiers and image segmentation with PR methods.

7.1.5. Study aims

The experiments conducted in this chapter are designed to compare the outputs of RF and SVM supervised global pixel-based classifiers against those obtained by RF and SVM classifiers that incorporate the spatial-contextual methods described above. My hypothesis is that using spatial-contextual methods will increase the accuracy and interpretability of spatially distributed classifications when compared to those obtained using global pixel-based methods. However, it is unclear which spatial-contextual classifiers are optimal for lithostratigraphic classification problems using geophysical data and whether there are differences in the abilities of RF and SVM to exploit contrasting spatial-contextual information. Target classes in this experiment represent generalised bedrock

lithostratigraphic units in a heavily forested region of western Tasmania, Australia. Several elements of this experiment, such as: limited T_a; high intraclass variability and interclass similarities; dense vegetation; and unconsolidated Quaternary sediments, represent challenging characteristics commonly encountered in supervised lithology classification applications. In addition, rather than solely evaluating classifiers based on classification accuracy, this experiment aims to identify the advantages and limitations of spatial-contextual methods with respect to computational cost and the generation of geologically plausible predictions.

7.2. Data

This section summarises the target classes, representing lithostratigraphic units, followed by a description of geophysical input variables and the pre-processing methods employed to prepare these variables.

7.2.1. Lithostratigraphy – classification target

The study region is located in western Tasmania and covers ~ 1000 km^2 (Figure 7.1).

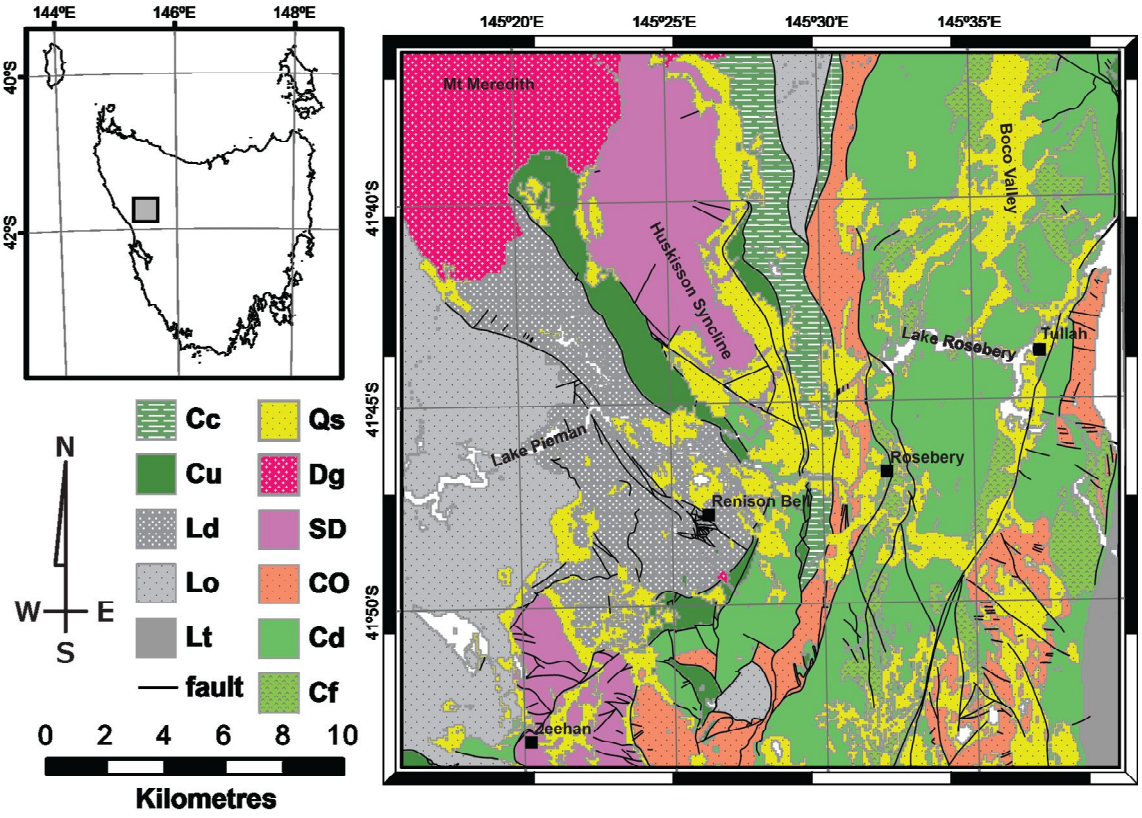

Figure 7.1 Study region location and generalised lithologic units, modified from Mineral Resources Tasmania (2011). Table 7.1 provides a summary of class descriptions and abbreviations.

Table 7.1 Summaries of lithostratigraphic classes. Classes are based on generalised/amalgamated geological units sourced from the 1:25,000 scale geological map of Tasmania (Mineral Resources Tasmania 2011).

Class	Period	Name	Description
Qs	Quaternary	Quaternary sediments	Unconsolidated sedimentary rocks
Dg	Devonian	Meredith Granite	Felsic intrusive rocks
SD	Siluro–Devonian	Eldon Group	Siliciclastic and calcareous sedimentary rocks
CO	Late Cambrian–Early Ordovician	Owen Group	Siliciclastic sedimentary rocks
Cd	Middle–Late Cambrian	Mt Read Volcanics	Clastic and volcaniclastic sedimentary rocks
Cf	Middle–Late Cambrian		Felsic volcanic and intrusive rocks
Cc	Early–Middle Cambrian	Cleveland Waratah Association	Mafic volcaniclastic sandstones
Cu	Early–Middle Cambrian	Ultramafics	Ultramafic-mafic intrusive and volcanic rocks
Ld	Late Neoproterozoic–Early Cambrian	Success Creek Group/Crimson Creek Formation	Siliciclastic and calcareous sedimentary rocks
Lo	Early Neoproterozoic	Oonah Formation	Metamorphosed siliciclastic sedimentary rocks
Lt	Mesoproterozoic	Tyennan	Metamorphosed siliciclastic sedimentary rocks

Lithological classes used to train and test spatial-contextual classifiers were obtained from the 1:25,000 digital geological map available from Mineral Resources Tasmania (2011). This map was compiled from geological reports based on 1:25,000, 1:50,000 and 1:63,360 scale mapping, new field mapping and interpretation of aerial photography and airborne geophysical data. Target classes were rasterised to 100 m pixel resolution using the polygon covering the pixel centre. Based on overall stratigraphic relationships and common geological materials, the original 1:25,000 scale lithostratigraphic units were generalised for training and prediction resulting in a total of 11 classes (Table 7.1).

The oldest unit in the study region is considered part of the Tyennan Region of central and southern Tasmania (Seymour & Calver 1995). This unit crops out in the southeast of the study region and contains a sequence of Mesoproterozoic–Early Neoproterozoic metapelitic and metaquartzitic sedimentary rocks (Turner 1989; Seymour & Calver 1995). This unit is in contact (unconformably) with overlying Palaeozoic rocks of the Dundas Region (see below) to the west. Early Neoproterozoic rocks of the Oonah Formation crop out in the west of the study region. The Oonah Formation is a structurally complex and highly variable sequence of dominantly quartzwacke turbiditic metasediments and minor volcaniclastic rocks. Lithologies within the Oonah Formation include muscovitic and

quartzitic sandstone, quartzite, turbiditic quartzwacke, volcaniclastic lithic wacke, pebble conglomerate, laminated siltstone, mudstone, carbonate rocks and basaltic lavas (Brown 1986; Turner 1989; Seymour & Calver 1995). The Oonah Formation is unconformably overlain by the Late Neoproterozoic–Early Cambrian Success Creek Group/Crimson Creek Formation (Brown 1986), although accessible outcrop within the study area indicates a faulted contact between these two units (Seymour & Calver 1995). Rocks within the Success Creek Group are dominated by quartz sandstone and laminated to thinly bedded mudstone and siliceous siltstone. The Success Creek Group is conformably overlain by the Crimson Creek Formation, which contains interbedded, volcaniclastic turbiditic wacke, siltstone and mudstone, with basalt lava horizons (Brown 1986; Brown 1989; Crawford & Berry 1992).

Early to Middle Cambrian mafic to ultramafic volcanic rocks and associated volcaniclastic rocks are thought to be a suite of spatially and genetically related allochthonous units. These allochthonous units were structurally emplaced as an east-dipping thrust-sheet over the Success Creek Group/Crimson Creek Formation (Rubenach 1974). Current tectonic models suggest these allochthonous units were obducted during the Early Cambrian Tyennan Orogeny (Berry & Crawford 1988; Crawford & Berry 1992; Turner 1998). Mafic to ultramafic rocks comprise sub-alkaline orthopyroxene-rich volcanic rocks with geochemical similarities to Ocean Floor Basalts (Brown 1989). Immediately east of the mafic to ultramafic volcanic rocks is a unit, considered a correlate of the Cleveland-Waratah association, compromising mafic volcaniclastic sandstone, siltstone, mudstone and chert with intercalated tholeiitic basalt flows (Mineral Resources Tasmania 2011; Seymour *et al.* 2013).

Following the emplacement of the allochthonous units in west Tasmania, a period of volcanism and sedimentation extended through the Middle and Late Cambrian (Seymour & Calver 1995). The Mt Read Volcanics were deposited in a submarine environment and comprise calc-alkaline felsic, intermediate and minor mafic volcanic rocks and host the majority of economic volcanic-hosted massive sulfide (VHMS) mineralisation in Tasmania (Corbett & Solomon 1989; Seymour *et al.* 2013). Included within the Mt Read Volcanics is an informally grouped sequence of volcano-sedimentary rocks. This sequence contains successions of interbedded tuffaceous sandstone, siltstone, shale, volcaniclastic conglomerate and breccia (Corbett & Solomon 1989; Seymour & Calver 1995). In this study, the Middle Cambrian volcano-sedimentary sequence, i.e. the Western Volcano-

Sedimentary Sequence (Mineral Resources Tasmania 2011), was merged with the volcano-sedimentary unit of the Mt Read Volcanics.

Unconformably overlying the Mt Read Volcanics is the Late Cambrian–Early Ordovician Owen Group, a sequence of siliciclastic sedimentary rocks (Banks & Baillie 1989; Seymour & Calver 1995). The Owen Group is dominated by coarse conglomerate featuring metaquartzite clasts derived from Tyennan Proterozoic sequences and includes non-marine (alluvial fan), through shallow marine, to deeper marine proximal turbidites (Noll & Hall 2005). Siluro–Devonian Eldon Group and minor Gordon Group rocks overlie Owen Group rocks. This unit outcrops within the Huskisson Syncline and in the southwest of the study area and comprises major sequences of quartz sandstone and siltstone with minor conglomerate horizons and limestone lenses (Banks & Baillie 1989).

Devonian intrusive rocks outcrop at Mt Meredith in the northwest of the study region (Brown 1986; Seymour & Calver 1995). Mt Meredith features a large composite granitoid body predominantly composed of two textural types. The southern part of the batholith features equigranular, fine-grained to medium-grained grey biotite adamellite. Whereas, the northern and eastern parts of the batholith features porphyritic biotite adamellite (Brown 1986). These intrusive rocks, collectively known as the Meredith Granite, display irregular contact geometries with adjacent country rocks and the associated contact aureole is up to 2.5 km wide (Groves *et al.* 1972).

Minor bodies of post-Carboniferous rocks (Jurassic dolerite, clastic sedimentary rocks of the Parmeener Supergroup and Cenozoic mafic volcanic/terrestrial sedimentary rocks) occur within the study region. These units, along with areas covered by water, were masked from the final set of classes due to their insignificant coverage. Basement rocks are in places obscured by Quaternary sediments including alluvium, swamp, glacial and glaciofluvial sediments (Augustinus & Colhoun 1986; Brown 1986).

7.2.2. Geophysical data – input variables

In this study, airborne geophysical data (available from Mineral Resources Tasmania at http://www.mrt.tas.gov.au) and Landsat ETM+ imagery (NASA 2002, available from the United States Geological Survey at http://eros.usgs.gov) were used as input variables. These data were transformed to a common coordinate system GDA94 zone 55 and

resampled to 100 m resolution. Prior to the spatial transformations indicated above, data specific pre-processing methods (summarised below) were applied.

7.2.2.1. Pre-processing

Total Magnetic Intensity (TMI) data were downward continued to the ground surface to enhance shallow magnetic anomalies (Telford *et al.* 1990). Downward continued TMI data were Reduced-to-Pole (RTP) using ERDAS ERMappper 7.2 and parameters (dec. = 13.1, inc. = 72.0) based on the Australian Geomagnetic Reference Field calculation (Australia 2012). RTP shifts the dipolar nature of the TMI anomalies as if the geomagnetic field is vertically orientated (Telford *et al.* 1990). A regional gradient, observed in the RTP magnetics data, was removed by subtracting a linear trend surface. Airborne Gamma-Ray Spectrometer (GRS) data (K, Th and U) were checked and corrected for negative values. In addition, the natural logarithm of the K/Th ratio was calculated and included as input. Landsat ETM+ band ratios were calculated corresponding to those described in Appendix C. Landsat ETM+ variables with Pearson's correlation coefficients > 0.85 associated with a large number of other Landsat ETM+ variables were removed. The pre-processing steps described above resulted in 11 variables for training and testing, i.e. DEM, RTP, K, Th, U, K/Th, Landsat 1, Landsat 4, Landsat 6, Landsat 3/5 and Landsat 5/7. These data were standardised to zero mean and unit variance to reduce the potential for scaling issues associated with the selection of samples using kNN, SVM and SOM algorithms.

7.3. Methods

This section summarises sampling methods used to obtain labelled samples (T) and subset these into T_a and T_b, for training and evaluation. In addition, methods used to construct spatial-contextual classifiers via the approaches presented in Sections 7.1.1–4 are described.

7.3.1. Data sampling

T_a were obtained using stratified random sampling. For each class, 100 T_a instances were sampled representing ~ 1 % of the total number of samples in the study region. This sampling procedure ensured that spatial-contextual classifiers were assessed with limited *a priori* knowledge of the spatial distribution of lithostratigraphic units (Grebby *et al.* 2011). Figure 7.2a provides an example of the spatial distributions of T_a. T_b comprised 10,000 randomly sampled instances independent of T_a. Figure 7.2b shows an example of the

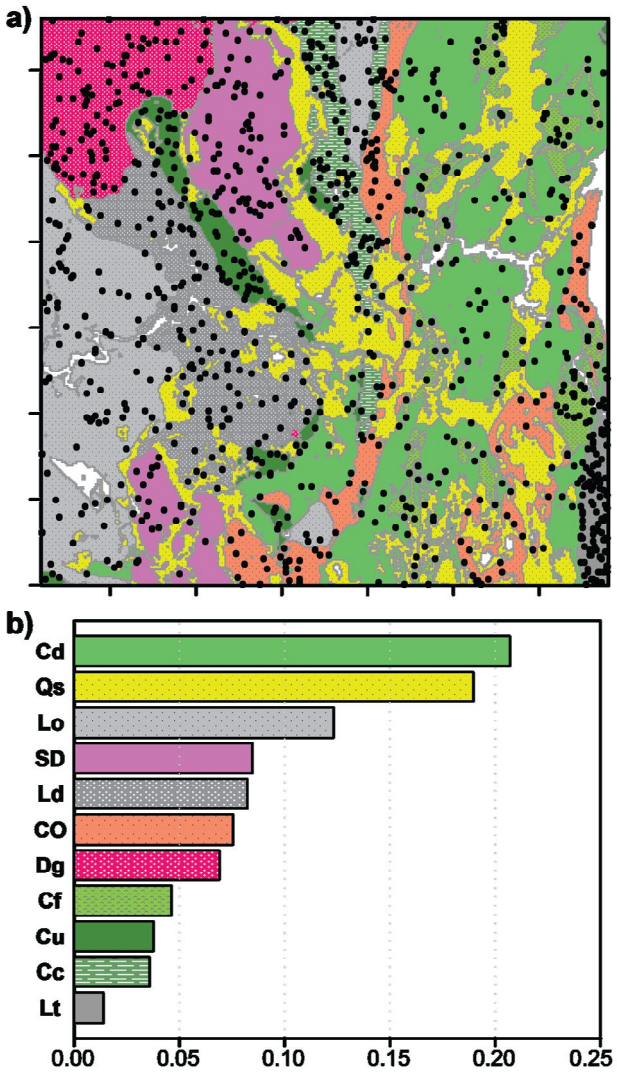

Figure 7.2 a) Example of T_a samples locations. Note Qs class was not included in T_a. b) Class proportions within T_b samples, this is approximately equivalent to total area covered by a given class. Note samples of Qs class not included in T_b for classifier evaluation. Table 7.1 provides a summary of class descriptions and abbreviations.

proportions of T_b samples for individual classes. These proportions are approximately equivalent to the area covered be each class within the study region. T_a and T_b were resampled 10 times each. Resampling was carried out to eliminate the possibility that classifications were influenced by the statistical and/or spatial distributions of randomly sampled data.

Samples of Qs, which cover ~ 20 % of the region under investigation, were omitted from T_a and T_b. Permian sedimentary rocks, Jurassic dolerite, Cenozoic basalt and water were omitted from T_a and T_b as these units either cover a small area of the study region and do

not have enough instances to sample adequately or, as for water, do not represent geological materials.

7.3.2. Global pixel-based classifiers

Global pixel-based supervised RF and SVM classifiers were trained and compared with spatial-contextual classifiers. Detailed descriptions of RF and SVM theory are provided in Sections 2.2.1.3–4. MLA parameters were selected using 10-fold cross-validation. An optimum minimum number of variables was obtained using the ranked-variable selection method described in Chapter 6 (Cracknell *et al.* 2014). RF global pixel-based classifiers were trained using 1000 trees and the 11 non-redundant geophysical variables were ranked using RF Gini Index measures of variable importance. In contrast, measures of variable importance calculated using Receiver Operating Curves (Provost & Fawcett 1997), as described in Chapter 4 (Cracknell & Reading 2014), were used to rank variables when training SVM global-pixel based classifiers.

7.3.3. Spatial-contextual classifiers

The methodologies employed to carry out supervised classification using spatial-contextual machine learning classifiers are summarised in this section. These methods are divided into those that are implemented during pre-processing, algorithm training and prediction post-processing stages of the MLA workflow.

7.3.3.1. Pre-processing

Focal operators with a 9 × 9 pixel neighbourhood were used to derive 1st and 2nd order spatial statistics. Edge effects were avoided by including a buffer of 1 km (10 pixels) surrounding the region under investigation. Five 1st order spatial statistics (mean, variance, coefficient of variation, skewness and kurtosis) and ten 2nd order spatial statistics, including nine GLCM texture derivatives (contrast, dissimilarity, homogeneity, angular second-moment, entropy, maximum probability, mean, variance and correlation) and fractal dimension, were derived from the 11 input variables documented in Section 7.2.2. GLCM texture derivatives were calculated from variables quantised to 64 grey levels (6-bit) as this has been shown to generate adequate results while reducing computational cost associated with high quantisation levels (Zortea *et al.* 2007). For each input variable, GLCM textures were derived for the four principle directions (0 °, 45 °, 90 ° and 135 °) and direction invariant texture (average of the principle directions). Pairwise grey-level comparisons were defined using a pixel offset of 2. Fractal dimension, an important

variable for lithology classification using airborne magnetics data, was calculated using the method described in Shankar (2009). The focal operators described above generated a total of 51 texture derivatives, henceforth called texture variables, for each individual input variable. Texture variables were standardised to zero mean and unit variance prior to MLA training and testing. In addition, the method described in Section 7.2.2.1 was used to remove correlated texture variables, resulting in a reduction in the total number of inputs from 561 to 166.

SOM was employed to automatically cluster the study region into 100 segments. SOM training was carried out using a 10 × 10 hexagonal grid with default parameters as defined in the *kohonen* package (Wehrens & Buydens 2007). A description of SOM theory and

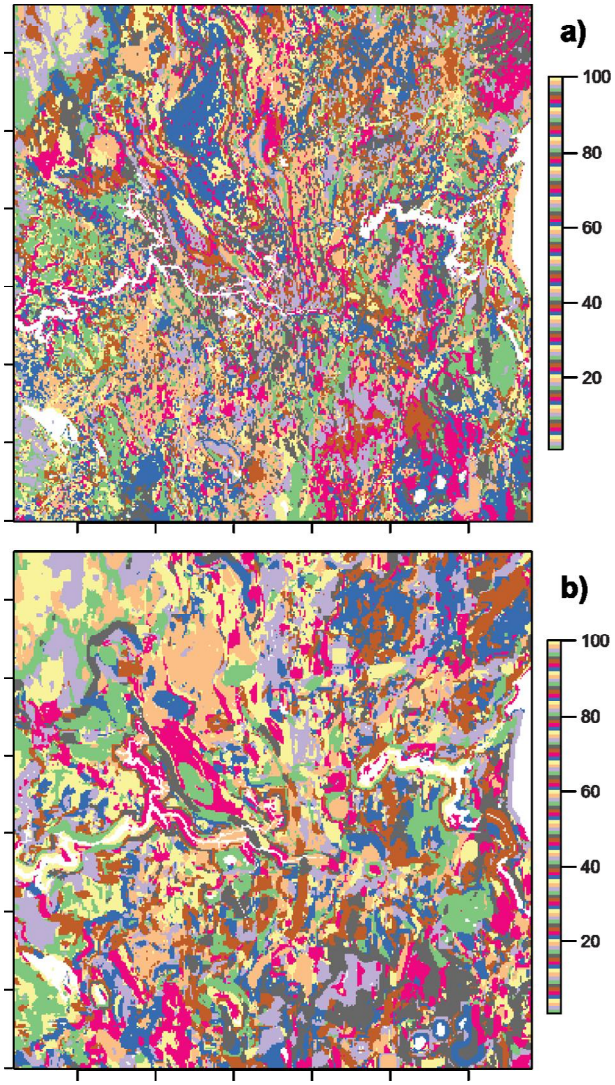

Figure 7.3 Segmented images derived using 100 SOM nodes for a) standard variables and b) texture variables. Note the colour ramp is repeated resulting in duplicated segment colours.

implementation is provided in Section 2.3.1.3. The resulting SOM segmented images for standard (Figure 7.3a) and texture (Figure 7.3b) variables highlight the effect of transforming variables using focal operators. The SOM segments derived texture variables are considerably larger and exhibit less irregular boundaries than the segments obtained using standard variables.

7.3.3.2. Algorithm training

In this study, I implement the fast-kNN (kNN-MLA) method, described in Blanzieri & Melgani (2008) and Zhang *et al.* (2006), to select neighbouring T_a samples in variable space. The kNN-MLA method trains a number of classification models equal to the number of samples in T_a. Classifiers were trained using k = 200 centred on each T_a sample, resulting in a total of 1000 local classifiers. To avoid cross-validation errors, resulting from not having all classes present within each cross-validation fold, 3-fold cross-validation was employed to optimise RF (*mtry*) and SVM (C) parameters. In addition, any class represented by < 10 neighbouring T_a was removed from the subset. Predictions were generated using the trained classification model associated with the closest (k = 1) T_a sample in variable space to the samples requiring classification. The ranked-variable selection method was not employed when training kNN-MLA classifiers as this would have led to prohibitively long processing times.

A new approach to the kNN-MLA method, which utilises SOM segments (clusters) to localise classifier training (SOM-kNN-MLA), was developed for this experiment. The SOM-kNN-MLA method trains a single classification model for every SOM segment. SOM node vectors, which represent the overall input variable characteristics of samples that fall within a given SOM node, are used to centre the T_a neighbourhood search. As with the design of the kNN-MLA method, k = 200 was used to subset neighbouring T_a samples for SOM-kNN-MLA classifiers. Samples requiring prediction were classified using the localised classifier associated with the SOM segment that contained the sample in question. In the SOM-kNN-MLA implementation, node-vectors representing the spectral properties of SOM segments, were used to define the centre of the k-nearest T_a search in variable space. Individual classifiers for each SOM segment were used to generate predictions for the T_b samples associated with respective SOM clusters.

7.3.3.3. Post-processing

An image segmentation and global pixel-based majority PR method (SOM-PR), similar to that described by Tarabalka *et al.* (2009), was implemented in this study. The SOM-PR method reclassifies categorical predictions by assigning labels to the pixels that fall within individual SOM segments the majority class predicted, within that segment, by a pixel-based classifier. In addition, I employed PR majority focal operators as an alternative means of introducing spatial context to MLA classifiers during post-processing. Majority focal operators assign the most abundant class in a given neighbourhood to the target (centre) pixel (Ricchetti 2000; Ghimire *et al.* 2010; Grebby *et al.* 2011). If two or more classes were found to be equally represented within a local neighbourhood the majority class associated with the target pixel was selected. Non-PR predictions were compared to those resulting from majority focal operators with neighbourhood dimensions of 3 × 3, 7 × 7 and 11 × 11 pixels.

7.3.4. Prediction evaluation

The effect of spatial dependencies complicates the interpretation and analysis of geospatial data (Burl *et al.* 1998). Random sampling provides a set of T_b that reliably indicates the success of supervised classifiers in regions close to T_a samples and in regions spatially disjoint to T_a samples. T_b performance statistics employed to compare classifications obtained by the methods described above were accuracy (and 95 % Confidence Intervals) and the kappa statistic. The kappa statistic is based on a confusion matrix and corrects accuracy estimates by accounting for random chance (Congalton & Green 1998). The spatial distributions of lithostratigraphic classifications and mismatch between the interpreted geological map and MLA classifications were assessed as a means of evaluating the geological plausibility of classifier outputs.

7.4. Results

Table 7.2 provides a detailed account of mean T_b accuracies (and 95 % Confidence Intervals) and kappa statistics for classifier predictions not reclassified using PR majority focal operators. All classifier mean kappa statistics are consistently ~ 0.05 less than mean accuracies. In addition, standard deviations from kappa values are equal or negligibly higher (0.002) than those obtained for accuracy. Mean 95 % Confidence Intervals are approximately ± 0.01 for all classifiers. Mean T_b accuracies range from < 0.60, for the

Table 7.2 Comparison T_b performance statistics for spatial-contextual classifiers. Mean and one standard deviation based on 10 spatially and statistically independent T_a and T_b resamples. Standard deviation of 95 % Confidence Intervals (CI) for all classifiers was << 0.001.

			Accuracy		95 % CI	kappa	
			mean	**st. dev.**	**mean**	**mean**	**st. dev.**
Standard Variables	**RF**	Global	0.625	0.009	0.011	0.575	0.010
		kNN-MLA	0.629	0.007	0.011	0.578	0.007
		SOM-kNN-MLA	0.624	0.008	0.011	0.573	0.008
		SOM-PR	0.652	0.013	0.011	0.606	0.014
	SVM	Global	0.581	0.014	0.011	0.521	0.014
		kNN-MLA	0.629	0.010	0.011	0.575	0.011
		SOM-kNN-MLA	0.627	0.010	0.011	0.573	0.010
		SOM-PR	0.608	0.013	0.011	0.551	0.013
Texture Variables	**RF**	Global	0.754	0.013	0.010	0.719	0.014
		kNN-MLA	0.696	0.016	0.010	0.652	0.018
		SOM-kNN-MLA	0.684	0.019	0.010	0.637	0.020
		SOM-PR	0.751	0.014	0.010	0.716	0.015
	SVM	Global	0.683	0.011	0.010	0.638	0.011
		kNN-MLA	0.671	0.019	0.010	0.620	0.019
		SOM-kNN-MLA	0.659	0.021	0.010	0.606	0.021
		SOM-PR	0.704	0.011	0.010	0.660	0.011

global pixel-based SVM classifier trained on standard variables, to > 0.75 for RF global pixel-based and SOM-PR classifiers trained on texture variables.

Figure 7.4 compares the statistical distributions of T_b accuracies for all classifiers. The use of standard variables result in 0.005 differences in mean accuracies between global pixel-based, kNN-MLA and SOM-kNN-MLA RF classifiers. RF SOM-PR obtained a ~ 0.03 increase in mean T_b accuracy compared to other RF classifiers trained on standard variables. The use of texture variables results in a > 0.10 increase in RF global classifier T_b accuracies when compared to the RF global classifiers trained on standard inputs. Global pixel-based and SOM-PR RF classifiers trained on texture variables achieve mean T_b accuracies > 0.05 than RF kNN-MLA approaches. When using texture variables, the best performing RF classifiers are the global and SOM-PR methods with considerable overlap between different resampled T_a and T_b.

SVM standard variable kNN-MLA and SOM-kNN-MLA classifiers achieved significantly higher T_b accuracies than the global (> 0.04) and SOM-PR SVM (> 0.02) classifiers. Considerable overlap in T_b accuracies is observed using SVM global and kNN-MLA (including SOM-kNN-MLA) classifiers trained on texture variables. The best performing SVM classifier was obtained using texture variables for training coupled with the SOM-PR

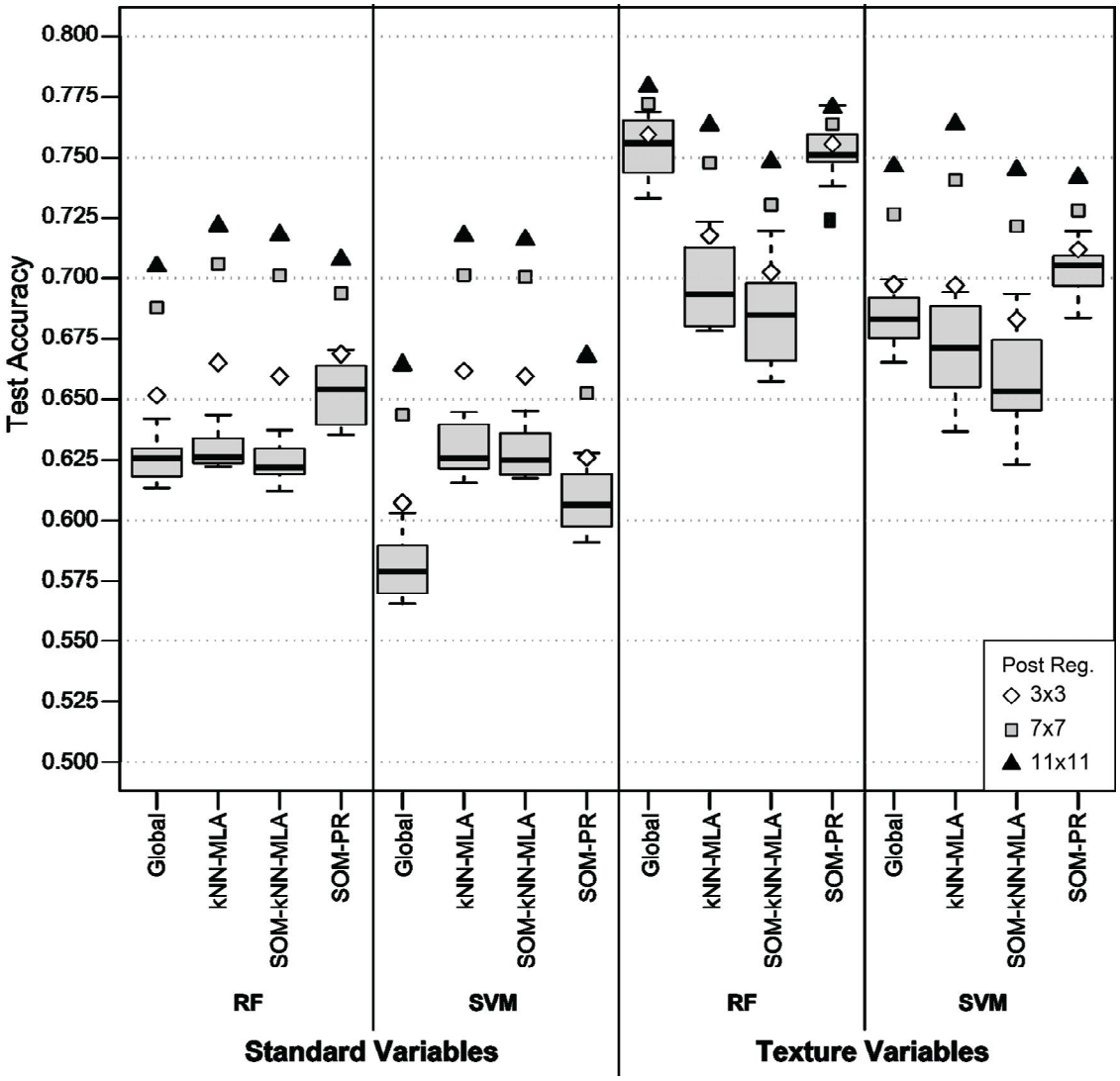

Figure 7.4 Comparisons of spatial-contextual classifier Tb accuracies. Boxplots indicate the distribution of classifier Tb accuracy obtained from 10 independent sets of Ta and Tb. Points (see legend) indicate mean Tb accuracy resulting from PR majority focal operator methods using different neighbourhood dimensions.

post-processing method. This method achieved a significant (> 0.02) increase in mean T_b accuracy compared to the other SVM classifiers utilising texture variables.

Substantial overlap between RF and SVM T_b accuracies generated using kNN-MLA and SOM-kNN-MLA, using standard or texture variables is observed. It takes considerably longer (× 10) to train kNN-MLA compared to SOM-kNN-MLA. Despite this, there are no significant differences in the mean T_b accuracies obtained by kNN-MLA and SOM-kNN-MLA classifiers trained on standard variables. When using texture variables both, RF and SVM implementations of kNN-MLA and SOM-kNN-MLA obtain mean T_b accuracies that fall within the ranges of standard deviations from the mean.

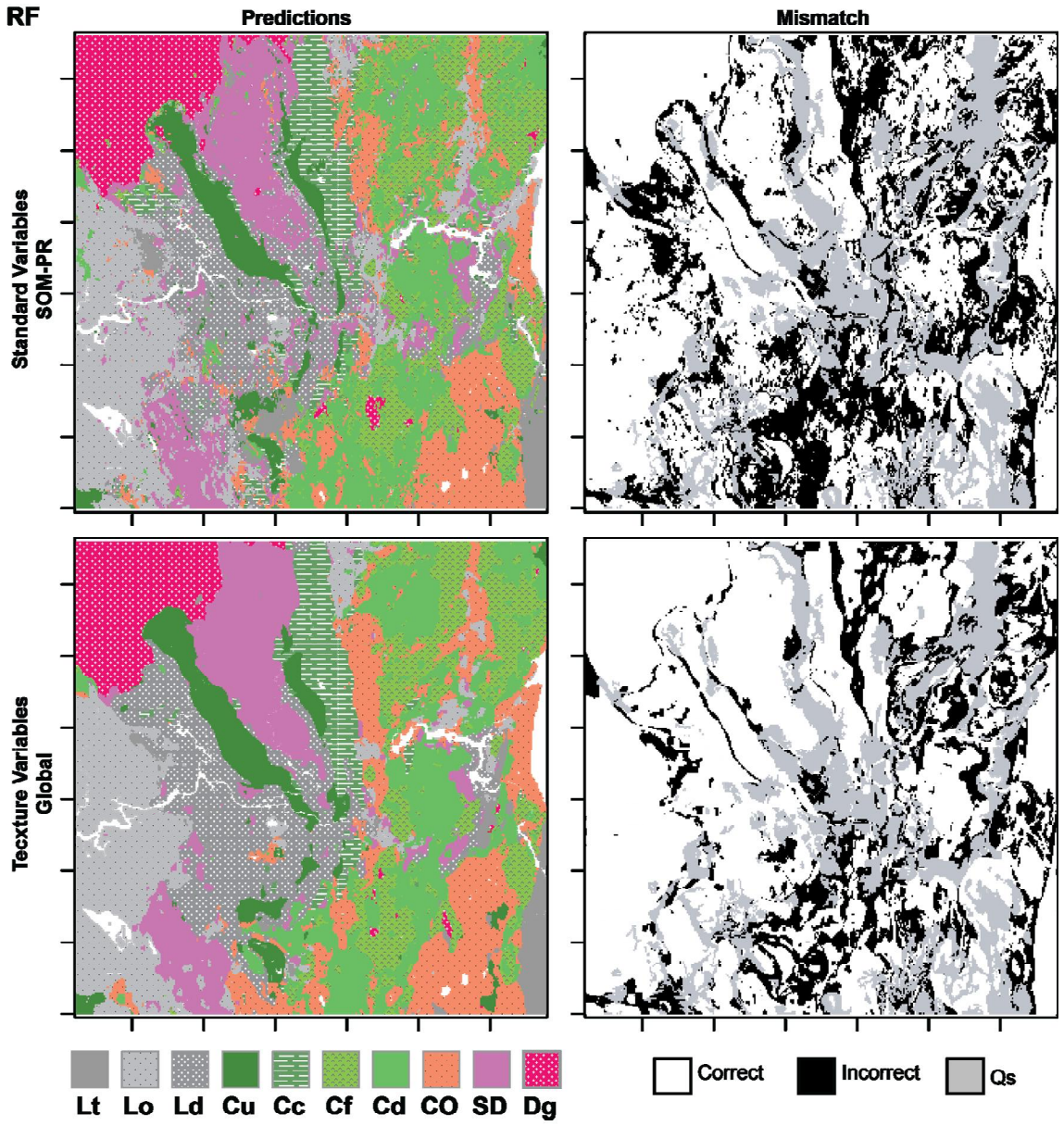

Figure 7.5 Selection of the best performing spatial-contextual RF classifiers.

Table 7.1 provides a summary of class descriptions and abbreviations.

Figure 7.5 compares the spatial distributions of classifications and the mismatch between these classifications and the interpreted geological map resulting from the most accurate RF classifiers trained using standard and texture variables. There is ~ 0.10 difference in mean T_b accuracies between these RF classifiers. The most striking difference in their classifications is observed in the degree to which errors (or correct predictions) are spatially contiguous using texture variables, i.e. these errors appear as connected bodies rather than classifications with a large degree of variation over small distances as observed using standard variables. A similar pattern emerges in Figure 7.6, which compares the best performing SVM classifiers utilising standard and texture variables. In this case, the SVM

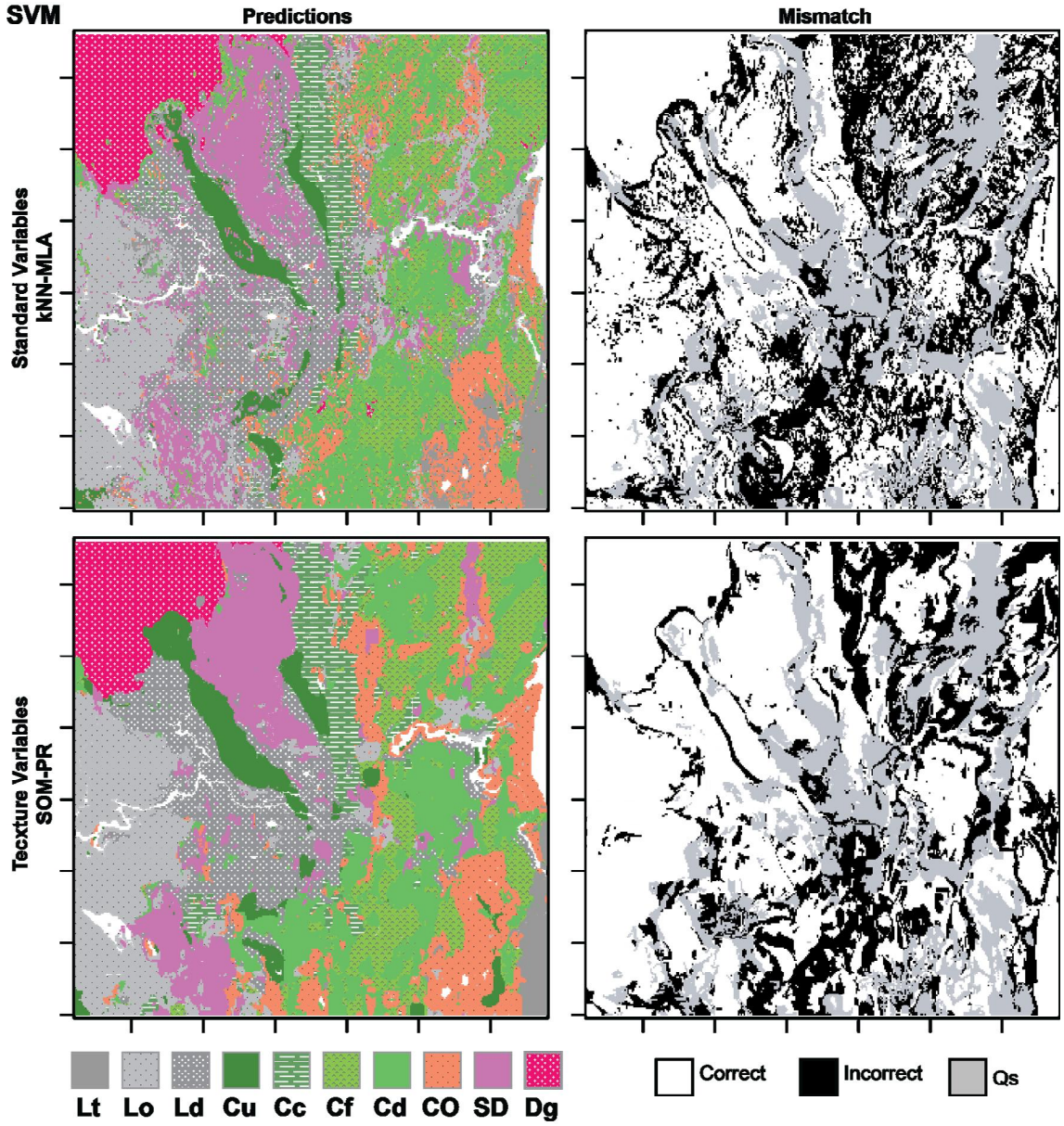

Figure 7.6 Example of best performing spatial-contextual SVM classifiers. Table 7.1 provides a summary of class descriptions and abbreviations.

SOM-PR classifier utilising texture variables achieves an ~ 0.075 increase in mean T_b accuracy over the SVM kNN-MLA classifier trained on standard variables.

Table 7.3 provides a summary of the differences observed in mean T_b accuracies obtained using PR majority focal operator. As the dimensions of PR focal operators increases there is a corresponding increase in the difference in mean T_b accuracies compared to predictions not subjected to PR focal operators. Classifiers trained using texture variables exhibit lower increases in mean T_b accuracies using 7 × 7 and 11 × 11 majority focal operators than classifiers trained on standard variables. The largest increase in mean T_b accuracies (~

*Table 7.3 Comparison of the difference between mean T_b accuracy obtained using PR majority focal operators of 3 × 3, 7 × 7 and 11 × 11 neighbourhood (pixel) dimensions and mean T_b accuracy resulting from predictions not utilising majority focal operators (see Table 7.2). Standard deviation of mean majority focal operator T_b accuracies is << 0.01 and mean of 95 % Confidence Intervals (CI) are ~ 0.01 ± 0.001 for all classifiers. * denotes PR T_b accuracy increase greater than the sum of one standard deviation and 95 % Confidence Intervals of mean T_b accuracy not using PR focal operators.*

			Difference in Mean Accuracy		
			3 × 3	**7 × 7**	**11 × 11**
Standard variables	**RF**	Global	0.026*	0.063*	0.080*
		kNN-MLA	0.036*	0.077*	0.093*
		SOM-kNN-MLA	0.036*	0.077*	0.094*
		SOM-PR	0.016	0.041*	0.055*
	SVM	Global	0.026*	0.062*	0.083*
		kNN-MLA	0.033*	0.073*	0.089*
		SOM-kNN-MLA	0.032*	0.073*	0.089*
		SOM-PR	0.018	0.045*	0.060*
Texture variables	**RF**	Global	0.006	0.019	0.026*
		kNN-MLA	0.021	0.052*	0.067*
		SOM-kNN-MLA	0.019	0.047*	0.064*
		SOM-PR	0.004	0.012	0.019
	SVM	Global	0.014	0.043*	0.063*
		kNN-MLA	0.026	0.069*	0.093*
		SOM-kNN-MLA	0.024	0.063*	0.086*
		SOM-PR	0.008	0.024*	0.038*

0.09) using majority focal operators are obtained by RF classifications trained on standard variables and SVM kNN-MLA classifiers trained on texture variables. There is a minor increase in mean T_b accuracies using 3 × 3 PR focal operators combined with RF and SVM classifiers trained on texture variables. In contrast, the greatest increases in mean T_b accuracies are observed using majority focal operators with dimensions larger than 3 × 3 pixels for kNN-MLA and SOM-kNN-MLA predictions.

Figure 7.7 compares RF and SVM classifier predictions trained on standard inputs and the effect on predictions resulting from the use of PR focal operators of different sizes. As the size of the PR focal operator neighbourhood increases there is a corresponding increase in spatially contiguous classifications with rounded (convex) outer edges. In addition, the original predictions must exhibit correct classifications within the local neighbourhood in order to reclassify predictions correctly as PR focal operators do not have the ability to distinguish potentially erroneous classifications. The impact of PR focal operators is reduced using texture variables because inputs have already undergone smoothing using

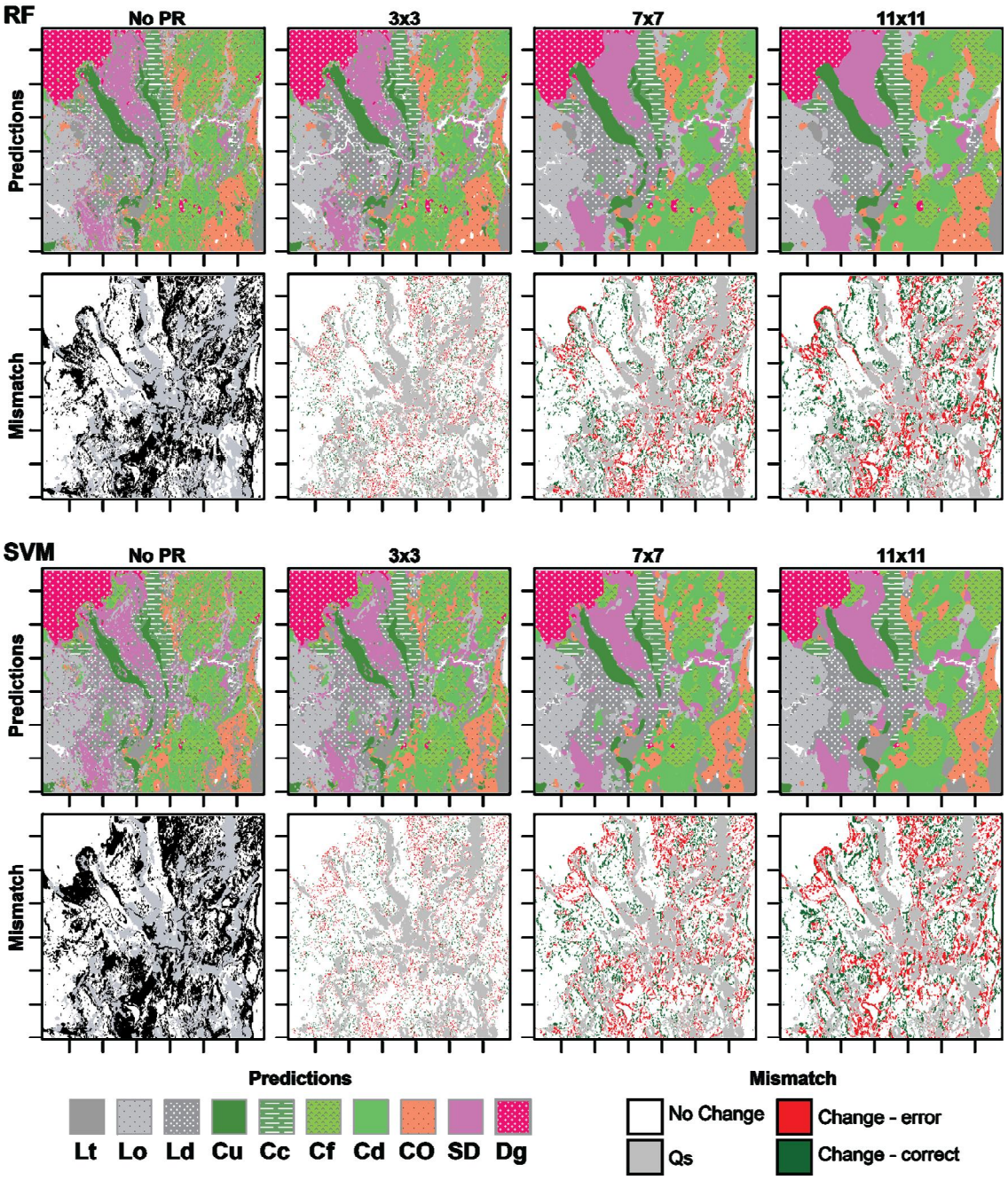

Figure 7.7 Example of RF and SVM classifications trained on standard variables. Mismatch images represent correct and incorrect classification as compared to the reference geological map used to train classifiers. PR mismatch images identify pixels that were reclassified using majority focal operators of different sizes. Colours indicate if the reclassification resulted in correct or incorrect reclassified predictions. Table 7.1 provides a summary of class descriptions and abbreviations.

local neighbourhoods. The results provided in Table 7.3 indicate that using PR focal operators of sizes much less than the size of focal operator neighbourhoods used to derive spatial statistics does not result in a significant increase in prediction accuracy.

Despite the significantly higher T_b accuracies compared to standard variables, classifiers utilising texture variables are unable to resolve classes representing lithostratigraphic units that are mapped as thin and discontinuous bodies. For example, the eastern body of the ultramafic unit, mapped as narrow region trending north-south, is not predicted by either RF or SVM using texture variables, although it does appear in predictions based on standard variables. The mapped extent of granite, in the vicinity of Mt Meredith, appears to have been classified with success by the majority of classifiers trialled in this study. Several small regions of granite are predicted in the region south of Rosebery by the majority of classifiers, especially RF SOM-PR using standard variables.

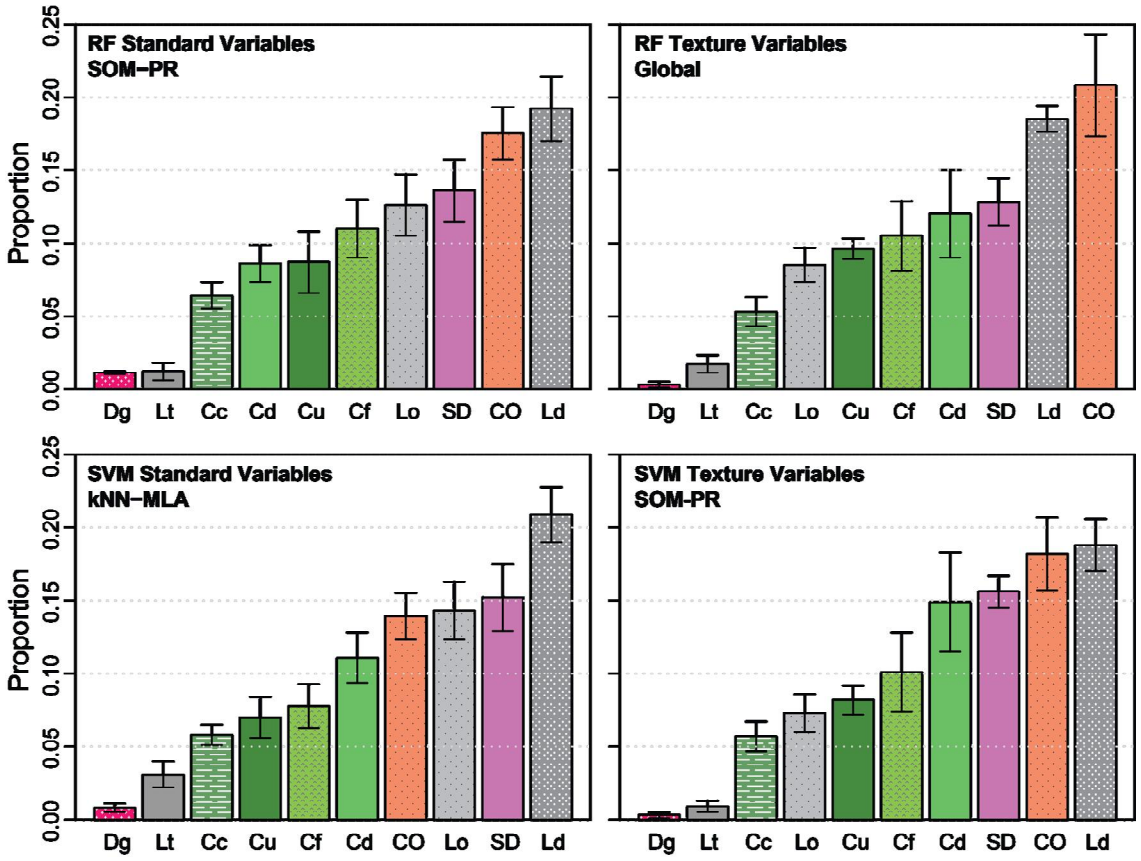

Figure 7.8 Comparisons of mean proportions (across 10 T_a and T_b resamples) of mapped Qs samples classified as classes present within T_a. Error bars indicate one standard deviation from the mean. Table 7.1 provides a summary of class descriptions and abbreviations.

Regions mapped as Quaternary sediments were not included in T_a. Quaternary sediment samples were, therefore, not assessed during the evaluation of classifiers. Figure 7.8 provides an indication of the proportion of individual classes generated by the most accurate spatial-contextual RF and SVM classifiers trained on standard and texture variables. These plots show Success Creek Group/Crimson Creek Formation, Owen Group and Eldon Group classes make up the majority of classifications for areas covered by Quaternary sediments. The bulk of Success Creek Group/Crimson Creek Formation and Eldon Group classifications in areas of Quaternary sediments occur on the southern limits of the Huskisson Syncline and east of Renison Bell. The bulk of Owen Group classifications cover spatially contiguous regions in the southeast of the study area. It appears that the banks of Lake Rosebery and the Boco Valley, in the northeast of the study region mapped as Quaternary sediments, contain the bulk of classification that are likely to be reflecting the presence of transported siliciclastic materials and not generating plausible classifications of bedrock lithostratigraphic units.

7.5. Discussion

The results of the experiments documented in this chapter are discussed in this section with particular focus on the use of spatial-contextual classifiers with respect to their advantages and limitations and issues identified surrounding spatial scale. This discussion leads into comments on the use of local geostatistical parameters as a means of optimising the scale at which spatial context is evaluated. Lastly, a brief interpretation of the geological relevance of classifications is provided.

7.5.1. Spatial-contextual classifiers compared

Global pixel-based classifiers trained on standard variables generated the lowest accuracies. The most accurate classifiers were trained using texture variables. Classifiers trained on texture variables were able to generate spatially homogeneous classifications not affected by high-frequency noise. However, the inability of classifiers trained on texture variables to resolve lithostratigraphic units expressed as narrow bodies suggests that the smoothing effect of focal operators inhibits the prediction of small scale features.

The use of kNN-MLA methods to train localised (in variable space) classification models significantly increases the T_b accuracies of SVM classifiers utilising standard variables. These results are to be expected and align with the findings of Segata & Blanzieri (2009)

and Segata & Blanzieri (2010). However, the resulting T_b accuracies of SVM kNN-MLA classifiers are equivalent to those obtained by global pixel-based RF classifiers trained on standard variables. In addition, there is no significant difference between RF global pixel-based classifier T_b accuracies and those observed for kNN-MLA RF classifiers. This reflects the adaptive-nearest neighbour architecture of RF classifiers, which encourages it to adjust the geometry of decision boundaries based on the distribution of local T_a in variable space (Lin & Jeon 2002; Hastie *et al.* 2009). Therefore, RF by default trains localised (in variable space) classifiers and appears to not require the use of methods that explicitly select T_a samples representing local variable characteristics.

The kNN-MLA method induces multiple localised classifiers and hence incurs significantly higher computational cost compared to training a single global pixel-based classifier. This computational cost is related to the fact that the number of classifiers is equal to the number of T_a samples. The SOM-kNN-MLA method, developed for this study, reduced computational cost by a factor of 10 and generated equivalent T_b accuracies to the kNN-MLA method. SOM was implemented by selecting an arbitrary number of seed-nodes from which to cluster samples. An optimal number of SOM nodes and appropriate SOM output map topology can be identified by searching different combinations of these SOM parameters and selecting those that result in minimum average quantisation and topological errors (Kiviluoto 1996; Uriarte & Martín 2005). In addition, utilising methods that select the most relevant variables for input into SOM clustering analysis is likely to generate better segmentation outcomes (Tarabalka *et al.* 2009).

PR methods are useful for generating spatially contiguous regions of a given class by eliminating high-frequency noise present in original classifications (Ricchetti 2000; Tarabalka *et al.* 2009; Ghimire *et al.* 2010). The SOM-PR method generated the highest T_b accuracies for RF using standard variables and SVM using texture variables. This method provided an efficient and effective means of limiting the effect of high-frequency noise on classifications. Ultimately the SOM-PR method results in higher overall T_b accuracies and more spatially homogeneous regions than global pixel-based classifiers (Tarabalka *et al.* 2009). However, as observed in the predictions generated using majority PR focal operators, the success of SOM-PR methods will be dependent on the bulk of samples being correctly classified within a given region. Incorporating prediction uncertainty measures, as described in Chapter 5 (Cracknell & Reading 2013), will provide a means of weighting predictions to reduce the effect of erroneous classifications on PR methods.

7.5.2. Issues of spatial scale

As discussed above, despite the higher overall mean T_b accuracies (and kappa) obtained for classifiers trained on texture variables, the size of focal operators is a limiting factor regarding the smallest features that can be predicted. Shankar (2009) suggests that the optimal neighbourhood dimensions should not be smaller than the smallest geological feature of interest. In contrast, the results presented in this chapter indicate that the neighbourhood dimensions of focal operators should not be larger than the smallest features of interest. This is because large neighbourhood dimensions can, in situations where the features of interest are relatively small, incorporate information from neighbouring features (Franklin *et al.* 1996).

The scale of the features under investigation is an important consideration when selecting appropriate neighbourhood dimensions for the derivation of focal operator 1st and 2nd order spatial statistics (Gahegan 2000). Nonetheless, Lloyd (2011) acknowledges that the selection of appropriate neighbourhood dimensions that adequately characterise the scale of target features is problematic. This is especially true when the scale of the target features is unknown and inconsistent between variables at different locations within the region under investigation.

Lloyd (2011) suggests that adaptive neighbourhoods can be identified by employing specific criterion, such as sampling densities, or by local optimisation procedures. The use of sampling densities is commonly encountered in spatial point pattern problems where adaptive kernel density estimation is required (Danese *et al.* 2008). However, using only the spatial distribution of T_a, i.e. the point pattern, is problematic in the types of applications described in this study because sampling density is unlikely to represent the spatial scale of the features under investigation. Alternatively, Franklin *et al.* (1996) proposed an automated optimisation method for estimating appropriate neighbourhood dimensions for image texture analysis. This procedure estimates the range parameter defined by the semi-variance curve for a given variable using a geostatistical approach. The range is taken to represent the maximum neighbourhood dimension (Franklin *et al.* 1996). However, this procedure is designed to estimate a fixed neighbourhood size for individual variables based on global spatial covariance parameters rather than adaptive neighbourhoods based on local spatial variability.

The problem of locally varying spatial scales with respect to the features of interest can be tackled using either multi-scale analysis or the estimation of non-stationary covariance parameters. Multi-scale textural derivatives can be generated for every sample and filtered, using variable selection methods, to identify the most relevant and meaningful data for input (Lu & Weng 2007; Guyon 2008; Pacifici *et al.* 2009). Alternatively, multi-scale textural derivatives can be randomly selected multiple times and combined via a weighted voting scheme using an ensemble of classifiers (Zortea *et al.* 2007; Gifford & Agah 2010). The estimation of non-stationary covariance parameters across the region of interest can be achieved using focal operators (Kanevski *et al.* 2004). As indicated by Franklin *et al.* (1996), the semi-variance range parameter can provide an estimate of the maximum neighbourhood size of a focal operator centred on a given sample. However, care must be taken to ensure that covariance parameters are estimated using an adequate number of samples, as too few samples will generate spurious results (Ramstein & Raffy 1989; Atkinson & Lewis 2000; Lloyd 2011). Furthermore, robust approaches to fitting trends to plots of semi-variance such as weighted least-squares is recommended (Hengl 2009). Due to the high computational cost of estimating multiple spatial covariance parameters, the use of parallel processing and/or image segmentation to reduce image space into a number of localised regions with equivalent characteristics is recommended.

The concepts discussed above regarding the estimation of local neighbourhoods with which to process variables prior to classifier training only considers isotropic covariance structures. Geological structures and hence, trends in the physical properties used to image geological features rarely vary equally in all directions, i.e. they are anisotropic (Boisvert & Deutsch 2011; Lloyd 2011). Anisotropy can be modelled locally such that the orientation of maximum continuity, i.e. the range, varies with location (Boisvert *et al.* 2009). Given the issues identified with the use of PR focal operators, both in terms of the scale of neighbourhoods use to identify majority classifications and the assumption of isotropy resulting in generally circular spatially homogenous regions, a weighted reclassification scheme could be employed that utilises estimates of anisotropic spatial variability to better represent geological features with specific structural trends.

7.5.3. Geological interpretations

The results presented in Figure 7.8 indicate that regions mapped as Quaternary sediments are dominated by classifications representing Success Creek Group/Crimson Creek

Formation, Owen Group, Eldon Group and Oonah Formation units. The spectral characteristics of these classes are likely to overlap with the characteristics of Quaternary sediments. This is because large amounts of siliciclastic materials, specifically derived from the Owen Group, constitute the bulk of Quaternary sediments in the study area (Augustinus & Colhoun 1986; Brown 1986). For example, thick glacial outwash sands and gravels in the Boco Valley (Augustinus & Colhoun 1986) obscure what is likely to be Mt Read Volcanics (McNeill & Corbett 1989). These Quaternary sediments were dominantly classified as either Owen Group or Oonah Formation. In contrast, the valleys surrounding Lake Rosebery are more likely to be classified as Eldon Group and (to the southeast) Success Creek Group/Crimson Creek Formation.

During the preparation of T_a and T_b, units representing the Western Volcano-Sedimentary Sequence and correlates of the Tyndall Group, within the Huskisson Syncline south of Lake Pieman (Mineral Resources Tasmania 2011), were merged with the volcaniclastic sediments of the Mt Read Volcanics. This unit comprises felsic to intermediate volcaniclastic sandstone, siltstone and chert-rich granule-pebble conglomerate rocks (Brown 1986). The majority of spatial contextual classifiers are classifying this region as containing a combination of siliciclastic and volcaniclastic materials, i.e. Success Creek Group/Crimson Creek Formation, rather than dominated solely by volcaniclastic sedimentary materials, i.e. volcaniclastic sediments of the Mt Read Volcanics.

A study conducted by Leaman and Richardson (1989) interpreted the presence of granitoid bodies at shallow depths (~ 1 km) south of Renison Bell and Rosebery from gravity data. Many of the RF spatial-contextual classifiers trialled in this study have predicted granite in the area described above.

7.6. Conclusions

The experiments detailed in this Chapter have evaluated new and pre-existing methods for introducing spatial-contextual information into machine learning algorithm (Random Forests and Support Vector Machines) supervised classifiers with respect to a challenging lithostratigraphy classification problem. Methods for including spatial-contextual information were separated into those implemented during (1) data pre-processing, (2) classifier training and (3) post-processing stages of the machine learning workflow. Spatial-contextual pre-processing generates measures of spatial similarity and dissimilarity

for input data using 1st and/or 2nd order spatial statistic focal operators. Spatial-contextual methods applied during classifier training utilise *k*-Nearest Neighbours sampling methods to identify and subset local training samples in variable space for the induction of multiple localised classifiers. Post-regularisation majority neighbourhood filters of different dimensions were used to reclassify predictions and generate spatially homogeneous classifications. Unsupervised clustering, implemented using Self-Organising Maps, was used to segment the area under investigation into spatially contiguous regions. Image segments were then combined with both *k*-Nearest Neighbours training data sampling and post-regularisation methods to speed up processing times and assist post-regularisation methods. These spatial-contextual methods and combinations thereof, were compared to global pixel-based classifiers, which unlike the methods described above treat individual samples as spatially and statistically independent and identically distributed.

The results of these experiments indicate that pre-processing data using focal operators (texture variables) significantly increase the classification accuracies of both Random Forests and Support Vector Machines as compared to using standard (non pre-processed) variables. The use of *k*-Nearest Neighbours methods for the selection of training data improves classification accuracies for Support Vector Machines. In contrast, no significant advantage in terms of classifier accuracy was observed for Random Forests using *k*-Nearest Neighbours methods to select training data. Moreover, combining textural data with *k*-Nearest Neighbours training sample selection decreased Random Forests classifier accuracy and did not result in any significant difference in the accuracy of classifiers trained using Support Vector Machines. Image segmentation combined with *k*-Nearest Neighbours sample selection significantly reduced processing times with no appreciable decrease in classifier accuracy using either standard or texture variables. Combining image segmentation with post-regularisation methods was more advantageous for Random Forests classifiers than for Support Vector Machines. Majority focal operator post-regularisation methods resulted in a substantial increase in classifier accuracy.

Despite the high overall classification accuracies obtained using texture variables and post-regularisation majority focal operators, a trade-off was observed between the abilities of classifiers to generate spatially homogeneous predictions and the degree to which fine-scale geological features could be resolved. This finding suggests the importance of identifying appropriate neighbourhood dimensions with which to derive texture variables and for reclassifying predictions.

CHAPTER 8 – SYNTHESIS AND DISCUSSION

The research documented in this thesis has investigated the practical advantages and limitations associated with the use of different machine learning algorithms for supervised classification. The applications detailed concentrate on integrating multivariate geological, geophysical and geochemical data to predict discrete 2D spatially varying classes representing lithological/lithostratigraphic units. This chapter initially synthesises my research findings regarding the use of machine learning algorithms for geospatial classification problems, followed by a discussion of the considerations required of geologists for their practical implementation. These insights are structured around the three stages of machine learning algorithm implementation: (1) data pre-processing; (2) classifier training; and (3) the evaluation of outputs. Particular attention is devoted to the appraisal of methods I have developed and employed for estimating prediction uncertainty and generating geological meaningful interpretations from the outputs of machine leaning algorithms. Finally, I discuss the significance of my research with respect to the emerging challenges faced by the wider geoscience community. In particular, I consider the applicability of this research in the rapidly advancing field of Big Data analysis and modelling.

8.1. Algorithms

In Chapters 4 (Cracknell & Reading 2014), 5 (Cracknell & Reading 2013), 6 (Cracknell *et al.* 2014) and 7, I detailed experiments designed to identify the advantages and limitations associated with the use of machine learning algorithms (MLAs) for practical geoscience mapping applications. Because geosciences data is often characterised by a limited number of direct observations, irreducible noise and a high degree of intraclass variability and interclass similarity, the choice of MLA must be appropriate to the context of these data. This section synthesises the insights and observations regarding the use of supervised classification and unsupervised clustering algorithms resulting from these experiments.

8.1.1. Supervised classification

I have evaluated and compared various MLAs corresponding to the five general machine learning strategies for supervised classification (Kotsiantis 2007): Naïve Bayes (NB) –

statistical learning algorithms; *k*-Nearest Neighbours (kNN) – instance-based learners; Random Forests (RF) – logic-based learners; Support Vector Machines (SVM); and Artificial Neural Networks (ANN) – Perceptrons. MLA comparisons were based on algorithm implementation, evaluation of their abilities to classify lithologies from geophysical remote sensing data, generating plausible geological maps and their response to variations in the spatial distribution of training data (T_a) and input variables.

8.1.1.1. Implementation

The results of experiments documented in Chapter 4 (Cracknell & Reading 2014) indicate that NB, SVM and ANN are substantially more sensitive to variations in classification model parameters than RF or kNN. Furthermore, RF and kNN essentially require the selection of a single parameter, making them relatively straight forward to optimise and train. The high sensitivity of NB, SVM and ANN classifiers requires the selection of appropriate model parameters that control the learning process. In addition, a large multi-dimensional parameter space must be searched in order to optimise the performance of SVM and ANN (Rohwer *et al.* 1994; Kotsiantis 2007; Oommen *et al.* 2008; Hsu *et al.* 2010). Where multiple model parameters require optimisation, the use of grid searches can be employed. Fine search grids can be computationally expensive, whereas, coarse search grids require discretisation and may ineffective at identifying optimal parameters for a given dataset (Guyon 2009). The use of parallelisation coupled nested parameter search methods can be used to reduce the computational cost of training. Alternatively, estimating specific parameter values, e.g. ó for SVM, eliminates the need to search multi-dimensional parameter space.

SVM incurred the highest computational cost during training, compared to the other algorithms trialled, despite the use of parallelisation and parameter estimation methods. The additional computational cost associated with the implementation of SVM is related to the quadratic optimisation of an objective function. This computational cost is directly related to the number of support vectors used to fit hyperplanes that separate classes and the number of classes and degree of spatial clustering in T_a. In contrast, kNN was found to be the fastest algorithm implemented during these experiments. However, when faced with datasets containing a large number of samples and many variables the computational cost of identifying neighbours and storing this information increases dramatically (Molina *et al.* 1994; Hastie *et al.* 2009). NB required substantially more processing time to generate predictions than to train classifiers, which may prohibit its use in situations where large

numbers of samples require classification. Table 4.3 (p. 83) indicates that SVM was unable to train classifiers more often than other algorithms and Table 5.3 (p. 109) shows that with spatially dispersed T_a containing small numbers of samples (~ 650) SVM was unable to train a classifier. This is due to fact that with high values of *C*, SVM could not converge on a stable solution and the cross-validation trial failed.

8.1.1.2. Decision structures

The decision structures induced by MLA supervised classifiers are governed by the inherent architectures of the respective strategies. Clear examples of the differences in the decision structures formulated by MLAs are presented in Figure 4.6 (p. 85). This figure plots the spatial distributions of MLA classifications given only spatial (X and Y) coordinates as input and provides a direct indication of the geometry of decision structures in 2D variable space.

When provided with a small number of spatially dispersed T_a samples and only spatial coordinates as input variables, SVM generated smooth, geologically plausible maps. This is distinct to the highly irregular maps, with approximately equivalent test data (T_b) accuracies, obtained using kNN and RF. Unlike other machine learning strategies, SVM algorithms define decision boundaries by maximising the margin between support vectors, which constitutes a minimisation of the geometric error in variable space (Burges 1998; Melgani & Bruzzone 2004; Ehret 2010), resulting in a situation where no local minima solutions exist (Masotti *et al.* 2006; Langer *et al.* 2009). The kernel function used to reproject variable space and the *C* parameter, which controls the degree to which support vectors are misclassified, allows SVM to fit arbitrarily complex (irregular) decision boundaries with distinct geometric characteristics that are a function of support vector locations in variable space. In contrast, the geometries of RF decision boundaries are orthogonal/parallel to the variable axes, whereas kNN decision boundaries explicitly honour Euclidian distances in variable space. Therefore, the orientation the axes in variable space governs the geometry of RF and kNN classifier decision boundaries.

8.1.1.3. Accuracy comparison

In Chapter 4 (Cracknell & Reading 2014), NB and ANN consistently obtained the lowest overall T_b accuracies of the MLAs trialled. When implementing NB it is assumed that input variables are conditionally independent for each class, which is unlikely to hold using geophysical data as input. Bayesian Networks (BN) employs a similar approach to NB,

although it does not require variables to be conditionally independent (Witten & Frank 2005) and may prove to generate more accurate results for classification problems that violate this assumption. However, for complex problems with a large number samples and variables it becomes computationally infeasible to calculate posterior probabilities using BN (Kotsiantis 2007). During ANN training, multiple local (error) minima can be identified resulting in overfitting (Ripley 1996; Burges 1998; van der Baan & Jutten 2000; Kotsiantis 2007; Hastie *et al.* 2009). Recently, Probabilistic (or Bayesian) Neural Networks (PNN) have been shown to circumvent the issues associated with ANN overfitting which results in suboptimal classifiers (Maiti & Tiwari 2010b). However, as PNN estimates network parameters using Bayesian probabilistic theory this approach is likely to add significant computational cost to classifier training.

The kNN algorithm is seen as a base level classifier for spatial remote sensing image classification (Brazdil & Henery 1994; Melgani & Bruzzone 2004; Hastie *et al.* 2009). This is because kNN implicitly exploits spatial dependencies within spatially distributed data by identifying a local classifier for each sample requiring prediction. However, kNN is susceptible to noisy data, especially when trained using low k values (Tan *et al.* 2006; Kotsiantis 2007; Hastie *et al.* 2009). This detrimental effect on kNN predictive capabilities is evident Figure 4.5 (p. 82) where a substantial decrease in T_b accuracies is observed when utilising geophysical input variables.

The research documented in this thesis indicates RF can be trained with minimal user intervention and limited T_a samples to generate highly accurate supervised classifiers from high-dimensional geophysical variables that contain irreducible noise. These findings are consistent with those reported in previously published research evaluating RF for a diverse range of remote sensing supervised classification problems (e.g., Ham *et al.* 2005; Pal 2005; Statnikov *et al.* 2008; Waske *et al.* 2009; Waske & Braun 2009; Ghimire *et al.* 2010; Duro *et al.* 2012; Ghimire *et al.* 2012; Gifford & Agah 2012; Waske *et al.* 2012). In situations where RF generated less accurate predictions than other MLAs, such as SVM, these differences in accuracy are minimal (Pal 2005; Statnikov *et al.* 2008; Waske *et al.* 2009). The results presented in Figure 4.5 (p. 82) indicated that as the degree of spatial clustering decreases RF generates substantially more accurate predictions than the other MLAs. Furthermore, the experimental results presented in Table 5.3 (p. 109) showed that RF achieved significantly ($p < 0.05$) higher mean overall accuracies compared to SVM given different numbers of T_a samples. In addition, Table 7.2 (p. 166) and Figure 7.4 (p.

167) highlights, when not employing multiple local spatial-contextual classifier (kNN-MLA) methods for subsetting T_a, RF obtains significantly ($p < 0.05$) higher mean T_b accuracies than SVM.

RF is an ensemble Decision Tree (DT) algorithm, it generates multiple classifiers and combines predictions using a majority vote in order to classify samples. The combination of classifiers prevents fitting to local minima and reduces the potential for classifier overfitting. In a recent study, Ghimire *et al.* (2012) evaluated bagging (Breiman 1996), boosting (Freund & Schapire 1996) and RF ensemble DT classifiers in the context of a land cover classification problem using Landsat ETM+ imagery. These ensemble classifiers were compared in terms of the number of T_a samples and variable levels of noise in T_a. The results of this study indicate that ensemble DT algorithms, other than RF, also show promise for the accurate classification of spatially distributed phenomena from noisy input variables.

8.1.1.4. Spatial-contextual classifiers

In Chapter 7, kNN was employed to identify neighbouring samples in variable space. This provided a subset of T_a samples for input into kNN-MLA classifiers using the method developed by Blanzieri & Melgani (2008). The results of my experiments indicate the kNN-MLA approach, which is designed to exploit spatial context, improves the accuracy of SVM classifiers but not those induced using RF. The architecture of RF is similar to that of an adaptive kNN classifier (Lin & Jeon 2002; Hastie *et al.* 2009), therefore, it is already implicitly utilising proximal T_a samples in variable space to the samples requiring prediction while also assigning higher weights to the most informative variables (Lin & Jeon 2002). As a result, no significant gains in RF classifier accuracy were observed using the kNN-MLA approach despite a significant increase in computational cost.

An alternative to subsetting T_a samples in variable space as a means of generating multiple local SVM classifiers is to adjust hyperplanes such that the support vectors are forced to acknowledge spatial context (Li *et al.* 2012). The computational cost of training these spatial-contextual SVM classifiers for a single iteration is only marginally more than for standard SVM. However, the spatial-contextual SVM requires multiple iterations thus significantly increasing the computation cost of implementing this algorithm.

8.1.1.5. Prediction uncertainty

In Chapter 5 (Cracknell & Reading 2013) I described and tested methods for estimating MLA prediction uncertainty. In this study, a modified version of *Variance* (Kohavi & Wolpert 1996) was used to obtain a scalar value describing the distribution of class membership probabilities. Other metrics have been used to derive uncertainty for machine learning classifications based on class membership probabilities, such as *Entropy*, 1 maximum probability and Best-versus-Second-Best (BvSB, Joshi *et al.* 2009; Loosvelt *et al.* 2012). Nonetheless, *Variance*, *Entropy* and 1 maximum probability uncertainty metrics behave equivalently (Appendix B, Loosvelt *et al.* 2012), while BvSB was shown to be useful for identifying samples requiring further clarification by users in an active learning application utilising SVM (Joshi *et al.* 2009). However, Loosvelt *et al.* (2012) suggests using metrics that evaluate the entire distribution of the class membership probabilities for geospatial applications, i.e. not BvSB or 1 maximum probability.

Modifications to the *Variance* metric result in standardised uncertainty values between 0 and 1 regardless of the number of classes in T_a. Standardising estimates of uncertainty provides a direct means of comparing the outputs of supervised classifiers with different numbers of potential classes. This is advantageous for the comparison of global pixel-based classifiers as the number of possible classes is governed by the number of classes within T_a. In situations where multiple classifiers are derived across a given spatial domain, each with different numbers of classes, non-standardised uncertainty values will be different. For instance, the multiple local classifiers obtained via the kNN-MLA method described in Chapter 7 are induced using subsets of T_a containing different numbers of unique classes. In this instance, using non-standardised measures of uncertainty can provide additional information on the relative complexity of classification models across the domain under investigation (Wellmann 2011; Wellmann & Regenauer-Lieb 2012). For example, the maximum possible value of *Entropy* is a function of the number of classes and thus the length of output class membership probabilities.

Two characteristics of geoscience data contribute to a high degree of mixed class types in the context of geophysical image classification: (1) the integration of data with different *support* (i.e. geometries and resolutions); and (2) the presence of intraclass variability and interclass similarities, within and between lithological units. These characteristics result in the presence of irreducible or deterministic noise within geophysical data (Scales & Snieder 1998). The results presented in Chapter 5 (Cracknell & Reading 2013) indicate

unequivocally that prediction uncertainties, derived from RF class membership probabilities, are a reliable indication of ambiguously classified T_b samples, i.e. multiple candidate classes with approximately equivalent probabilities. High SVM uncertainty was not correlated with erroneous classifications. This is because SVM class membership probabilities are a function of the distance of a given sample to the hyperplane boundary in variable space. This in turn is linked to the choice of kernel, in this case a Gaussian Radial Basis Function (RBF), used to transform variable space so that non-linear SVM decision boundaries are possible (Niaf *et al.* 2011). In contrast, RF utilises the normalised proportion of votes for each class obtained by the ensemble of DT classifiers (Hastie *et al.* 2009). The fundamental differences between SVM and RF, in terms of class membership probability estimation described above, result in the contrasting uncertainty outputs observed for these MLAs.

8.1.2. Unsupervised clustering

The main focus of this thesis has been to utilise data representing observations to generate geological maps via supervised classification approaches. Despite this, the Self-Organising Maps (SOM) unsupervised clustering algorithm was used as a complementary tool in the machine learning workflow. In Chapter 6 (Cracknell *et al.* 2014), SOM was employed to identify geologically meaningful and spatially contiguous lithological sub-classes within volcanic units predicted by RF. In Chapter 7, SOM was employed to segment the original spatial domain into contiguous regions with similar characteristics. In addition, SOM clusters were used to generate localised classifiers in conjunction with the kNN-MLA method. I demonstrated that by segmenting image space using SOM a significant reduction in the computational cost of training local kNN-MLAs was achieved without compromising predictive accuracy. SOM derived segments were also used during post-processing as an alternative post-regularisation (PR) method to majority focal operators (Tarabalka *et al.* 2009).

The implementation of SOM in the experiments documented in Chapters 6 (Cracknell *et al.* 2014) and 7 utilised default parameters for: the number of iterations; initial search radius and radius decrease over iterations; and percentage adjustment of seed-node properties. Parameters for the resolution (number of seed-nodes) and dimensions of the output 2D SOM map were defined manually. The choice of manually defined SOM parameters was governed by the expected number of discrete clusters in the resulting

segmentation outputs. Segmentation via SOM obtained spatially contiguous groups of samples associated with different nodes. This suggests clusters were produced that adequately described the natural groupings of samples in variable space (Fraser & Dickson 2007). Despite this, the manual methods employed to define SOM 2D map resolution and dimensions may be suboptimal. The process of defining an optimal number of SOM seed-nodes and appropriate SOM output map dimensions can be semi-automated by searching different combinations of SOM parameters and selecting those that result in minimum average quantisation and topological errors (Kiviluoto 1996; Uriarte & Martín 2005). This approach is similar to that described by Paasche & Eberle (2009), which utilises the Xie-Beni index (Xie & Beni 1991). The Xie-Beni index provides an indication of cluster separation by assessing the distributions of samples associated with a given cluster with respect to overall cluster characteristics. Thus, the Xie-Beni index can be used to identify an optimal number of natural groups with which to cluster variable space.

8.2. Applications

In this section I discuss the practical considerations required to implement MLAs with respect to the three stages of MLA implementation methodology employed throughout this thesis: (1) data pre-processing; (2) classifier training; and (3) the evaluation of outputs. Specific focus is placed on synthesising the knowledge gleaned from my research with regard to the challenges faced by users when employing MLA supervised classifiers for geospatial inference problems. These challenges relate to: non-Gaussian data distributions; the presence of irreducible noise or irrelevant information within input variables; high intraclass variability and interclass similarities; variable spatial distributions of limited numbers of labelled samples (T) with which to train and test classifiers; MLA parameter selection; variables that exhibit spatial heterogeneity and dependencies; robust statistical assessment of spatially distributed outputs; and the formulation of meaningful interpretations from MLA decision structures.

8.2.1. Data pre-processing

Pre-processing input variables is a fundamental aspect of generating accurate and interpretable MLA supervised classifiers for practical applications. Sourcing and preparing appropriate data for a given application requires a certain degree of intervention by expert users. This is especially true where variable specific transforms or combinations of

variables are required to enhance relevant signals. This section closes with a discussion of the automated methods employed in this thesis for the identification of relevant variables.

8.2.1.1. Data preparation

The practical implementation of MLAs starts by identifying and sourcing data for a given application. For spatial inference problems, data quality, availability, coverage and their support (type of observation, scale/resolution and sampling density) are the limiting factors that place restriction on which data can be used as input into the inference process. This is because most MLAs require a full set of variables for all samples (both T_a and T_b) in order to train classifiers and evaluate outputs (Witten & Frank 2005; Shi & Liu 2011).

For spatially continuous variables, interpolation is an established approach to assigning values to missing data, e.g. inverse distance weighting, minimum curvature, etc. (Burrough & McDonnell 1998). The specific interpolation method employed for a given application will be dependent on the spatial distribution and resolution of observations. The majority of data used in this thesis was supplied as spatially continuous fields, i.e. it was already interpolated from discrete irregularly spaced observations. Whereas, in Chapter 6 (Cracknell *et al.* 2014) I used kriging to generate images of the spatial distribution of soil geochemical assay data.

The majority of MLAs can handle a range of data types such as numerical, categorical and ordinal in both discrete and continuous forms (Michie *et al.* 1994b). This is useful in situations where one or more input variables are derived from the analysis of other data, for example, samples assigned to discrete groups based on the outputs of unsupervised clustering methods (e.g., Stepinski & Bue 2006; Bue & Stepinski 2007).

Normalisation may not be an issue for MLAs that can emulate arbitrary statistical distributions, e.g. ANN and DT classifiers. However, standardisation (scaling) is required to mitigate the effects of the curse-of-dimensionality especially for kNN and SVM classifiers (Brazdil & Henery 1994; Friedman 1997; Wettschereck *et al.* 1997; Hastie *et al.* 2009; Hsu *et al.* 2010). In order to efficiently compare different MLAs I have found it sensible to implement all classifiers using identical transformed variables. This eliminated additional complexities when formulating statistical and spatial comparisons.

8.2.1.2. Variable extraction

Variable extraction is the process by which pre-processed variables are transformed such that the component signals represented by these inputs are enhanced for the intended application. In its simplest form, variable extraction includes the combination of two variables, for example, as a ratio (Guyon 2008). The choice of variable combinations and transforms is often defined by the expert users understanding and knowledge of the intended application. Specific corrections and transformations of input variables are usually required to enhance signals representing the classification target. For example, in this thesis Total Magnetic Intensity (TMI) data were Reduced-to-Pole (RTP), downward continued and regional residuals calculated during pre-processing as a means of enhancing the signal of near-surface geological materials (Telford *et al.* 1990). In contrast, Landsat ETM+ data, sourced with Level 1 processing applied, were not corrected for atmospheric effects. This is because the regions under investigation were contained within a single Landsat image (consisting of multiple bands) and classifications were generated using T_a and T_b from this single image space. However, when attempting to classify multiple Landsat images using classifiers trained on data from spatially disparate regions with contrasting atmospheric conditions, atmospheric corrections are essential (Song *et al.* 2001). Additional variable extraction methods implemented within this thesis utilised combinations of variables as ratios, e.g. Gamma-Ray Spectrometry (GRS) and Landsat ETM+ data. Ratios were based on information from previous studies (e.g., Durning *et al.* 1998; Kusky & Ramadan 2002; Inzana *et al.* 2003; Boettinger *et al.* 2008; Mshiu 2011) that identified particular combinations of GRS and Landsat ETM+ data as beneficial for the discrimination of lithologies and minerals.

Chapter 7 showed that the use of spatial-contextual variable extraction methods lead to significant improvements in RF and SVM classifier accuracies. This is because the variable transform methods employed implicitly characterised the spatial heterogeneity and dependency of spatially distributed variables. An alternative method of variable extraction is Principle Component Analysis (PCA). Despite the common use of PCA in multivariate geological/geoscience inference problems (e.g., Granath 1988; Gelfort 2006; Grebby *et al.* 2011; Yu *et al.* 2012), PCA was not investigated in any detail as it transforms the input variables such that the contributions of a given variable to the interactions with the target classes is difficult to interpret. This is especially true when PCA is used to construct linear combinations of all variables (Jolliffe 2002). Simplistic interpretation can be misleading

unless careful thought has gone into the choice of input variables and whether to transform them prior to analysis. In many cases the user must employ their knowledge of the problem to interpret the interactions between the outputs of a mathematically derived linear function of all the original variables and the outputs derived from these in order to decipher their interactions in complex multiclass inference problems (Jolliffe 2002).

8.2.1.3. Variable selection

The selection of appropriate input variables for a given application is an aspect of the machine learning workflow that requires careful consideration. This is because most if not all MLAs are impacted by the inclusion of redundant and/or irrelevant data (Wettschereck *et al.* 1997; Guyon 2008; Hastie *et al.* 2009). A common approach is to utilise expert knowledge of the inputs that are likely to be relevant to the intended application (Henery 1994a; Burl *et al.* 1998; Lu & Weng 2007) or based decisions on individual sensor characteristics, e.g. spatial, spectral and temporal resolutions.

In this thesis, I have documented an automated approach to variable selection using two complementary methods. Firstly, the removal of highly correlated or redundant inputs via methods that employ Pearson's correlation coefficients was used to select non-redundant data. The selection of non-redundant data has, in addition to generating more reliable predictions for algorithms that are challenged by the inclusion of correlated data (Domingos & Pazzani 1997; Witten & Frank 2005), the benefit of reducing computational cost and results in a reduced variable space from which to clarify and interpret data interactions. Secondly, I used a ranked-variable selection method to reduce input variables to a minimum number relevant to the intended application. The motivation for selecting a minimum number of relevant variables is to improve prediction accuracy, reduce computational cost and aid user understanding of the learning process.

8.2.2. Classifier training

In this section, I summarise and discuss aspects of MLA training that affect the classification outcomes. First and foremost is the need to investigate and understand the characteristics of what is known about the problem, i.e. T, in order to make appropriate decisions on the division of these data into T_a and T_b subsets. Secondly, during classifier induction decisions must be made regarding MLA parameter selection and, in the context of spatial-contextual classifiers, the selection of appropriate parameters when generating

local classifiers. Thirdly, the use of MLA classification post-processing provides users with a range of options that have the potential to improve classifier outcomes.

8.2.2.1. Training and test data

Supervised classification requires T, which representing what is known about the target phenomena, in order to train and evaluate the classification generated by MLAs. The division of these data into independent sets of T_a and T_b must be carefully considered. This is because both T_a and T_b should adequately represent the population distributions of the phenomena under investigation. It is important to have an adequate number of representative T, where an adequate number, in this case, will be defined by the relevance of the input variables to the target classes, the number of classes and separability of these classes (van der Baan & Jutten 2000; Link & Blundell 2003). Unlike previous studies (e.g., Ham *et al.* 2005; Gelfort 2006; Oommen *et al.* 2008; Waske *et al.* 2009; Song *et al.* 2012), which examined the influence of the number of samples in $\hat{O}_a$ on overall T_b accuracy, the experiments conducted in Chapter 4 (Cracknell & Reading 2014) investigated the role of spatial clustering on MLA performance. These experiments identified that the use of spatially dispersed (scattered) T_a avoided the issues associated with T_a spatial dependencies and provided a representative population of individual class variable characteristics to MLAs.

In order to reliably compare different algorithms and remove the possibility of training bias classifiers due to specific characteristics of a given T_a, random sampling was employed to generate T_a for the experiments conducted in this thesis. An alternative to controlling the number of samples for each class in T_a, a situation that is unlikely in real-world applications, is to use weighted cost functions (Tan *et al.* 2006). Weighted cost functions force classifiers to concentrate on correctly classifying classes with higher weights and are used to overcome imbalanced T_a (see Appendix A). However, user input is required to assign weights to classes.

8.2.2.2. Classifier induction

As discussed in Section 8.1.1.1, MLAs require the selection of algorithm specific parameters in order to minimise the generalisation error on T_a. The optimisation of model parameters is commonly conducted using methods such as cross-validation or bootstrapping (Efron 1983; Kohavi 1995; Breiman 1996; Witten & Frank 2005; Hastie *et al.* 2009). The bulk of the experiments documented in this thesis used 10-fold cross-

validation to selected model parameters as it generates stable estimates of classifier performance (Kohavi 1995; Guyon 2008). However, when selecting optimal model parameters for the kNN-MLA spatial-contextual classifiers employed in Chapter 7, 3-fold cross-validation was used. This was because low number of samples in T_a for individual local classifiers created a situation where not all classes in T_a were present in each of the cross-validation folds. This resulted in failed cross-validation trials and the inability to train classifiers.

Figure 4.5 (p. 82) shows that in situations where T_a are highly spatially clustered, classifications generated by MLAs were sensitive to variations in T_a. This is evident from the high degree of variance observed from the multiple trials for highly spatially clustered T_a. Conversely, as the degree of spatial clustering of T_a decreased a corresponding improvement and decrease variability in T_b accuracies was observed. Minimising the variability in T_b accuracies between classifiers trained on different T_a implies that increasingly stable classifiers, which capture the overall statistical distributions of data, are being trained. This has the benefit of reducing the difference between cross-validation accuracy estimates and T_b accuracies as indicated in Table 4.3 (p. 83). These results show a reduction in classifier overfitting with a decrease in the spatial clustering of T_a. These findings have implications for collecting observations for real-world applications where the intention is to utilise MLAs to generate first-pass models of the spatial distribution of classes representing natural phenomena.

8.2.2.3.Classification post-processing

In this thesis I have documented the use of two classification post-processing methods, both of which were shown to result in substantial improvements in accuracy. In Chapter 5 (Cracknell & Reading 2013), uncertainly thresholds obtained from the assessment of T_b classifications were used to identify the majority of erroneous predictions. In Chapter 7, MLA classifications were reclassified using spatial-contextual PR methods.

My research into the estimation of geospatial supervised classifier uncertainty, unlike other similar studies (e.g., Goodchild *et al.* 1994; Zhu 1997; Brown 1998; van der Wel *et al.* 1998; Loosvelt *et al.* 2012), demonstrated that RF prediction uncertainty could be used to identify erroneous classifications. The majority of incorrect predictions were selected via numeric thresholds derived from the analysis of T_b misclassifications. Figure 5.6 (p. 110) indicated that this novel use of classifier uncertainty was especially beneficial for

improving the accuracy of RF classifiers trained on small numbers of samples. Furthermore, I was able to show that uncertain predictions were spatially coincident with lithology contacts and areas intense deformation.

Classification post-processing, using spatial-contextual PR methods has been used to improve classifier accuracies in many remote sensing supervised classification applications (e.g., Ricchetti 2000; Tarabalka *et al.* 2009; Ghimire *et al.* 2010; Grebby *et al.* 2011; Li *et al.* 2012). PR methods are based on the assumption that spatially distributed features, such as land cover classes or lithologies, occur as distinct and homogeneous regions. In this thesis, two spatial-contextual PR methods were employed to improve classification accuracy. The first method (SOM-PR) utilised SOM nodes to group pixels into spatially contiguous segments. These segments were then used to reclassify global pixel-based classifications such that the majority class within individual segments was deemed to be the correct class. The SOM-PR method was observed to generate slightly lower variability in classifier T_b accuracies obtained from different T_a for both RF and SVM classifiers trained on variables, pre-processed using focal operators, which represent spatial-contextual information. The second PR approach employed in Chapter 7 used majority focal operators. A majority focal operator assigns the most commonly occurring class within a given spatial neighbourhood surrounding a pixel to that pixel. Significant increases in mean T_b accuracies using majority focal operators with dimensions larger than 7 × 7 pixels were observed, due in part to an increase in the spatial homogeneity of classifications at large scales. Despite this, a trade-off was observed with respect to the scale of features that could be adequately represented with large PR majority focal operator dimensions.

8.2.3. Evaluation and interpretation

In this section I initially discuss the use of statistical methods for assessing MLA outputs. Several methods have been employed during the course of my research to statistically evaluate supervised classifications generated by MLA supervised classifiers. Many statistical approaches, such as overall accuracy, recall, precision and kappa require independent T_b with known class labels. In real-world applications, T_b is a subset of T. However, estimating the likely performance of classifiers on new data and generating predictions for these data is often not the sole aim of implementing MLAs. In many

situations, gaining knowledge of the interactions between variables and the phenomena under investigation is an important element of the inference process (Feyyad 1996).

There are two approaches to interpreting or gaining a conceptual understanding of the machine learning process: (1) interrogating statistical decision structures constructed by an algorithm during training (Cutler *et al.* 2007; De'ath 2007; Guyon 2008; Hastie *et al.* 2009); and (2) processing geospatial machine learning outputs using complementary methods such as spatial analysis and unsupervised clustering (Park *et al.* 2003; Toumani 2003; Zhou & Chen 2005; Dohm *et al.* 2007; Zammit *et al.* 2007; Langer *et al.* 2009; Kraut & Wettergreen 2010). Both approaches require methods for analysing and visualising the relationships between variables, classes and/or auxiliary data using statistical and spatial methods.

8.2.3.1. Statistical evaluation

Accuracy, which provides an estimation of the probability that a classifier will correctly predict an unknown sample, is a standard measure of supervised classification performance (Congalton & Green 1998; Lu & Weng 2007). Additional measures of classifier performance documented and utilised in this thesis include confusion or error matrices, the kappa statistic and individual class recall and precision rates. Nonetheless, these measures of classifier performance require independent (to T_a) T_b with a minimum number of samples to estimate with confidence (Foody 2009; Hastie *et al.* 2009). Classification accuracy coupled with exact 95 % Confidence Intervals provide an indication of what the differences in accuracy of classifiers might be allowing for stronger conclusions to be drawn about the significance of classifier performance (Foody 2009).

Alternative methods to overall accuracy for assessing classifier outputs involve the use of confusion matrices, which represent classification results in tabular form. Confusion matrices offer a means of understanding the distribution of multiclass classifications within T_b (Kuncheva 2004). In Chapter 6 (Cracknell *et al.* 2014), the confusion matrix presented in Table 6.4 (p. 138) employs a novel approach to assessing lithological classifications that structures matrix rows and columns in stratigraphic order. This confusion matrix configuration not only provides an indication of the distribution of classifications across classes but offers a means of identifying geologically meaningful features within the results. This approach clearly showed that the bulk of misclassifications were concentrated

within the Que–Hellyer Volcanics (QHV) and that the bulk of misclassification occurred between stratigraphically and thus spatially proximal lithological units.

The kappa statistic is based on the counts derived from a confusion matrix but corrects overall accuracy for agreement that occurs by chance (Landis & Koch 1977; Congalton & Green 1998). The results documented in Chapters 4 (Cracknell & Reading 2014) and 7 indicate that the kappa statistic, while consistently lower than overall accuracies, does not show any differences in the rankings of MLAs to those based on overall accuracy. As with the kappa statistic, individual class recall and precision rates are derived from a confusion matrix. Recall and precision are good statistical measures for identifying if particular classes are over or under represented in the distribution of correct classifications. These measures of classifier performance were used to provide a means of identifying shifts in the distribution of individual class predictions in Chapters 5 (Cracknell & Reading 2013) and 6 (Cracknell *et al.* 2014).

8.2.3.2. Interrogating decision structures

In the previous section I discussed methods used in this thesis to statistically evaluate and compare MLA classifications. However, merely obtaining an indication of the probability that classifications will be accurate does not imply that we have gained understanding that facilitates an interpretation of the causal relationships governing the phenomena under investigation (Henery 1994a; Ripley 1996; Shmueli 2010). In the context of machine learning geological mapping, simply generating an accurate map or model does not provide geologists with a conceptual understanding of formative geological processes. The attainment of a conceptual understanding of geological phenomena offers geologists the opportunity to utilise confidently a geological map for making decisions (Brodaric *et al.* 2004). In this thesis, meaningful geological interpretations have been formulated from multivariate decision structures embedded within trained MLAs in two ways: (1) assessing the relative importance of selected variables; and (2) interrogating the interactions between variables and target classes.

Due to the difficulties traditionally associated with interpreting non-linear “black box” algorithms, such as RF and ANN (Breiman 2001; Miller & Han 2001; Hastie *et al.* 2009), the process of variable selection offers a simple means of reducing the number of inputs to a few that can be interpreted directly (Guyon 2008; Hastie *et al.* 2009). Throughout this thesis the removal of correlated variables was employed to reduce the number of input

variables to a minimum. Plots of relative normalised variable importance (see Figure 4.3, p. 80) were used to visualise the relevance of individual variables to lithology classes. This approach to visualising the relative importance of variables was replicated in Chapter 6 (Cracknell *et al.* 2014). In this case, in-built RF methods for plotting the relative importance of variables via the Gini Index were used. The RF method for estimating variable importance is preferred when executing ranked-variable selection methods because additional processing is not required. In contrast, the univariate pair-wise class combinations method implemented in Chapter 4 (Cracknell & Reading 2014) for the other MLAs incurs significant additional computational cost.

Plotting the relative importance of variables with respect to all classes, while giving a good indication of the variables that are contributing the most to classifier training, does not provide users with the ability to identify the interactions between specific variables and individual classes. In contrast, partial dependence plots (Friedman 2001) provide a graphical representation of the relative influence of variables on the prediction probability of individual classes after the (average) effects of the other variables have been accounted for (De'ath 2007). Class membership probabilities generated by any algorithm can be used to construct partial dependence plots (Hastie *et al.* 2009). However, partial dependence plotting functions are embedded in the R programming *randomForest* package (Liaw & Wiener 2002) allowing users to assess relative measure of the predictive strength (importance) of individual variables with ease.

Partial dependence plots have been used to interrogate the interactions between variables and classes based on the RF classifier trained in Chapter 6 (Cracknell *et al.* 2014). Figure 8.1 shows RF partial dependence plots of four soil geochemical (Zr, Ti, Cr and Cu) variables and their relative univariate effect on the prediction of QHV units in the Hellyer–Mt Charter region. These variables were chosen because they represent important geochemical variables defined by RF and/or they are seen as immobile elements useful for discriminating between primary igneous rocks in deeply weathered terranes (Hallberg 1984). These partial dependence plots indicate that the majority of QHV units are more likely to be predicted in regions represented by relatively low Zr and high Ti values. An exception to this is seen in: andesite dominated units (both feldspar-phyric footwall andesite and hangingwall andesite units), which display high partial dependence at ~ 200 ppm Zr; and dacite, which is more likely to be predicted in areas of Zr values < 200 ppm and > 250 ppm. This coupled with higher dacite partial dependence associated with Ti

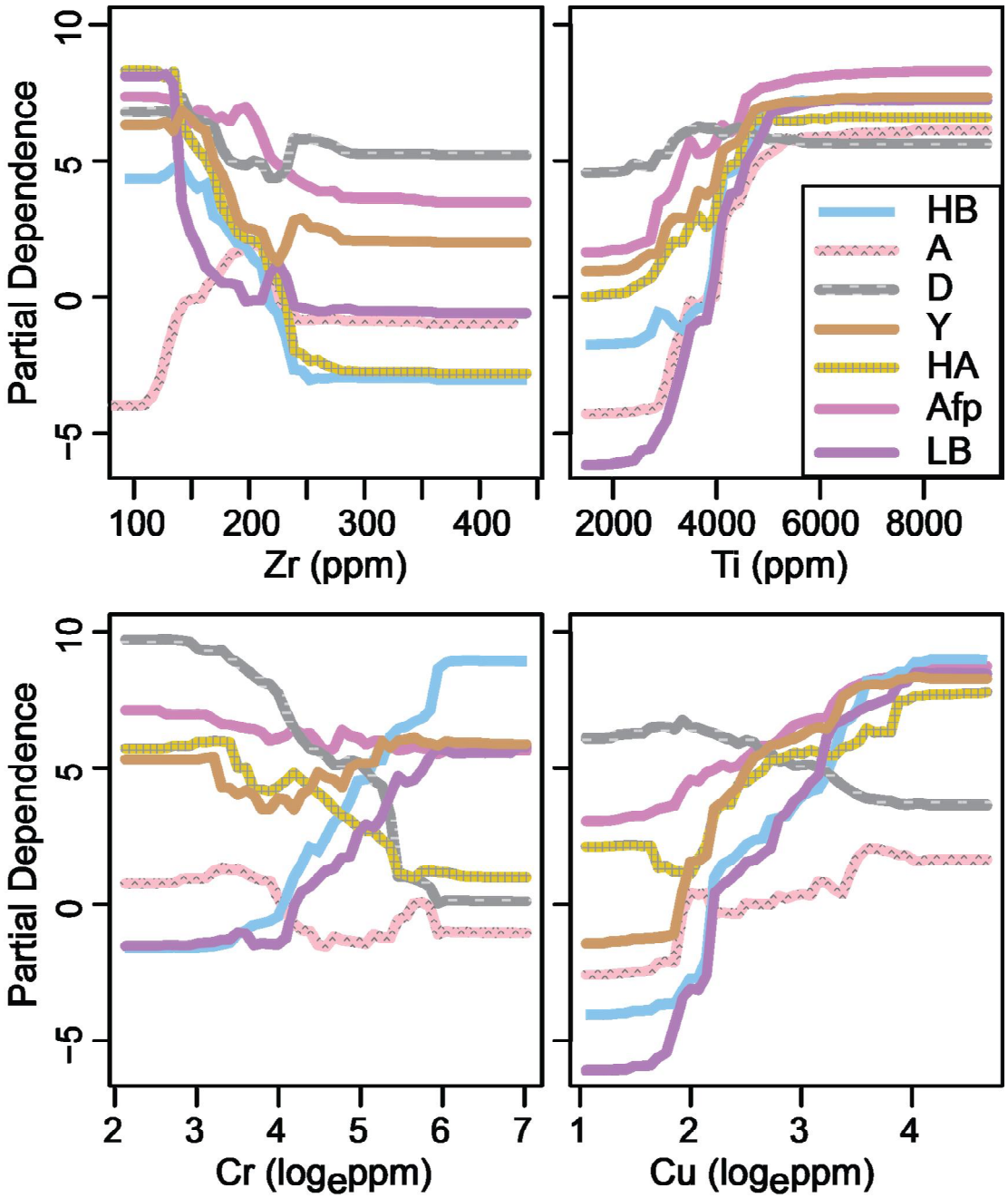

Figure 8.1 RF partial dependence plots showing the relative influence of variables on the prediction presence of QHV units. The y-axes represent the logit probability of prediction. Note that the trend of individual class partial dependence is important not the absolute differences between classes. See Table 6.1 (p. 126) for a description of class abbreviations.

values ranging from 3000–5000 ppm suggests two major groups within the dacite unit. The Cr partial dependence plot clearly indicates the Lower Basalts and Hellyer Basalts are associated with elevated Cr levels. In contrast, dacite and strongly altered rock exhibit strong relationships to low Cr values. Both the polymictic volcaniclastic unit and overlying andesite unit display associations with low and high levels of Cr. High Cu concentrations

are associated with QHV footwall, VHMS altered units and the Hellyer Basalt. These observations highlight geochemical contrasts representing primary compositional and influence of alteration on the likelihood of predictions.

The examples provided above explore interactions between input variables and target classes as a means of explaining or interpreting the geological significance of the decision structures induced by RF during classifier training. However, as Shmueli (2010) acknowledges, there is often a distinction between accurate predictive models and informative explanatory models. For example, in Chapter 7 the most accurate MLA classifiers were those that utilised input variables transformed to represent spatial context. If one was to interrogate the interactions between these transformed inputs and target classes the significance of these learned relationships may not be immediately obvious. This observation requires the user to carefully consider the motivations for employing MLAs to a specific problem within the context of desired outcomes. If the aim of the modelling exercise is to generate accurate predictions of geological phenomena in disparate spatial domains to that of T_a, the use of transformed variables is highly recommended. In contrast, if the intention is to test a hypothesis in order to gain understanding of geological phenomena within a local spatial domain, the use of original variables will provide better opportunities to achieve this research outcome.

8.2.3.3. Complementary interpretation

Throughout this thesis, comparisons between interpreted geological maps and the spatial distributions of classifications have been crucial for the comparison and assessment of MLAs. Plotting the spatial distribution of classifications provides a direct and easily interpretable indication of the efficacy of MLAs to generate geologically plausible predictions. In addition, plots of the spatial distributions of misclassifications, or mismatch between interpreted geological maps used to obtain T_a for MLA comparisons, offer a clear indication of the distributions of potential classification errors. These plots, similar to Figure 7 published by Yu *et al.* (2012, p. 236), can be used in conjunction with individual class recall and precision rates to identify and interpret differences between MLA classification models. However, in real-world situations where T_a represent field observations and little is known about the geology of a particular domain, comparison with existing maps is infeasible.

Spatial models of RF prediction uncertainty represent a robust means of evaluating the plausibility of classifications representing geological materials without the need for comparisons against pre-existing geological maps. My research into the estimation of RF prediction uncertainty provides geologists with a practical basis for the analysis and visualisation of uncertainties associated with geological maps inferred from geophysical data, which in turn can be used to communicate the associated limitations and complexities (Wellmann & Regenauer-Lieb 2012). This is clearly demonstrated in Figure 5.4 (p. 107) where an uncertainty map is used to assist interpretation of the validity of RF classifications and identify regions of the resulting geological map that are likely to be in error. Furthermore, uncertainties provide additional data for future applications. For example, in Chapters 5 (Cracknell & Reading 2013) and 6 (Cracknell *et al.* 2014), I discuss the use of RF uncertainty as means of optimising and directing future field work. In this instance, data collection can be focussed in regions that offer the opportunity to better constrain geological maps by identifying locations that will contribute informative T_a to subsequent MLA classifiers.

Remotely sensed data that images the Earth's surface or near-surface may, in some cases, be subject to the influence of physically and chemically weathered bedrock materials, i.e. regolith (Scott & Pain 2008). In these situations the characteristics of bedrock materials are likely to be concealed by *in situ* and/or transported geological materials. The focus of the research documented in Chapters 4–7 was to investigate the use of MLAs to map bedrock lithologies in regions where weathering process are assumed to be constant. Where weathering processes are not uniform across a given region it is likely that MLA predictions of bedrock materials from remote sensing data will be affected by weathered materials.

Geological materials are rarely homogeneous in their composition. In Chapter 6 (Cracknell *et al.* 2014) I used the SOM unsupervised clustering algorithm to identify spatially contiguous but distinct sub-classes within the samples predicted as QHV andesite and basalt volcanic units. Plots of the relative contributions of key variables to SOM sub-classes, in conjunction with their spatial distributions provided meaningful interpretations of compositional differences present within these volcanic units. Compositional differences were interpreted to relate to differences in primary volcanic compositions due to contrasting phases of eruption and differences in the degree and type (footwall or hangingwall) of VHMS alteration.

8.3. Extended research implications

In this section I initially discuss the implications of my research in the context of an integrated workflow using the R programming language. This is followed by a summary of the broad range of geoscience applications that can be addressed using the methods developed in this thesis. This section closes with comments on the implications of my research to the emerging need to improve scientific research and decision making outcomes via the analysis and modelling of Big Data.

8.3.1. Integrated workflow using R

The methods and analysis presented throughout this thesis have been implemented using the R programming language (available from Comprehensive R Archive Network – http://cran.r-project.org/). R is a stable and reliable open source statistical software platform that provides users with the opportunity to develop and test code. Being open source, R is freely available and can be distributed and/or modified without fear of breaking licensing agreements. R also offers users the ability to adapt a wide range of existing packages and functions and code new functions for specific purposes. R provides functions and packages that can input and output a wide range of data formats used by other software. Moreover, R may be used to implement and control other software during the inference process. For example, R can be integrated with Matlab™, Python, ArcGIS™, GRASS and Database Management Systems such as Microsoft Access™ and SQL Servers.

One of the key aspects of R that has been exploited throughout this thesis is the ability to integrate the MLA workflow, i.e. data pre-processing, MLA training and output evaluation, into a single software environment. This allows for the inference process to be efficient and amenable to a diverse range of practical applications (Burl *et al.* 1998). Table 8.1 summarises the R code and scripts provided in the (digital) Appendix F, which were developed during in the course of my research and used to conduct experiments documented in Chapters 4 (Cracknell & Reading 2014), 5 (Cracknell & Reading 2013), 6 (Cracknell *et al.* 2014) and 7. These scripts include functions and routines for: spatial data input and analysis; image analysis and interpretation; machine learning training and parameter optimisation; and map assessment and visualisation. This code can be freely distributed, thus, providing researchers with the opportunity to implement the machine learning methods documented in this thesis.

Table 8.1 Summaries of the R code and scripts provided in (digital) Appendix F.

File /Script	Description
raster_functions.r	Collection of functions to process, analyse and visualise raster data.
spatial_projections.r	Collection of Coordinate Reference System (CRS) proj4text objects. Settings for other proj4text projections for and CRS can be sourced from http://spatialreference.org/.
machine_learning_functions.r	Collection of function to sample labelled data, select variables and visualise machine learning outputs.
spatial_sampling_functions.r	Collection of functions to conduct spatial sampling, e.g. stratified sampling and sampling using contrasting spatial geometries.
spatial_statistics_functions.r	Functions to calculate 1st and 2nd order spatial statistics for raster objects. Includes Grey Level Co-occurrence Matrices (Haralick *et al.*, 1973) with options for controlling direction (orientation), step size (offset) and width (dimensions).
BrokenHill_comparison.r	Run experiments and analyses as described in Chapter 4. Includes pre-processing, spatial sampling, MLA training and evaluation for comparisons.
BrokenHill_uncertainty.r	Run experiments and analyses as described in Chapter 5. Includes methods for obtaining, analysing and visualising prediction uncertainty.
Hellyer.r	Run experiments and analyses as described in Chapter 6. Includes variable selection methods and routines for implementing and visualising SOM unsupervised clustering outputs.
Rosebery_spatial-context.r	Run experiments and analyses as described in Chapter 7. Includes routines for the derivation and selection of texture variables, kNN-MLA training sample selection and classification post-regulariation.

8.3.2. Wider geoscience applications

The computer-assisted geological mapping methodology and workflow demonstrated in this thesis provide a practical and efficient means of: integrating geoscientific data, i.e. geological, geophysical and geochemical variables, for geological mapping; estimating uncertainty; and gaining knowledge from statistical and spatial structures present within geoscientific data. Experiments into the use of MLAs for geological mapping applications were conducted in three contrasting geological terranes at a variety of scales and resolutions. These experiments highlight the wide applicability of the described methods to remote sensing geological mapping problems in a diverse range of geological terranes.

Recently, the Australian Academy of Science initiated a vision for joint partnerships between research institutes and industry which aims to promote advances in mineral exploration under cover (UNCOVER 2012). This initiative acknowledges the challenges faced by mineral explorers for targeting ore bodies concealed under cover rocks. Fundamental aspects of improving ore deposit targeting outcomes include the need to

integrate and analyse disparate geoscientific data collected at multiple scales and characterise and manage uncertainties derived from the inherent limitations of data and incomplete evidence (UNCOVER 2012). Given appropriate input variables, such as those derived from gravity, magnetics, electromagnetic and seismic observations, the techniques developed in this thesis have the potential to deliver real-world outcomes in the search for deep, buried ore systems. In addition, the robust estimates of uncertainty documented in Chapter 5 (Cracknell & Reading 2013) will enable better characterisation of the confidence that can be assigned to the spatial elements of geological models.

The identification of ore deposit footprints within cover sequences and the characterisation of sub-class representing alteration vectors have not been explored fully in the context of targeting buried ore deposits (UNCOVER 2012). In Chapter 6 (Cracknell *et al.* 2014), I demonstrated that geologically meaningful variability within lithological units due to differences in primary composition and alteration styles could be identified using unsupervised clustering. These methods can be easily transferred to the task of characterising contrasting depositional environments and contrasting magmatic phases, which offer an opportunity to interpret geological histories and tectonic environments.

Geological maps offer geologists the opportunity to identify and characterise geological formative processes (Leverington & Moon 2012). These maps are also a fundamental base layer of information for wide range of problems and applications (Thomas 2004) such as: ore deposit prospecting and modelling; tectonic reconstructions, geohazard risk assessment and engineering applications; geomorphological and hydrological studies; and ecological research. Given that geological maps are a crucial element in many decision making endeavours, it is imperative that these maps are accurate or contain robust indications of model uncertainty. This is because unidentified errors in geological maps will propagate through subsequent analysis. Nonetheless, limited field observations used to constrain often subjective geological interpretations means that error and uncertainty are an ever present component of geological maps. Despite this, geological maps rarely provide an indication of the uncertainties of the underlying interpretations (Bárdossy & Fodor 2001; Lindsay *et al.* 2013). This is because it is difficult to assign robust estimates of confidence to subjective interpretations made during the production of most geological maps. My research indicates that using MLAs, such as RF, to critically evaluate pre-existing interpretations produces more accurate geological maps. These updated maps, in conjunction with robust measures of uncertainty, provide analysts from a wide range of

disciplines with information that will aid interpretations based on the results of subsequent analyses.

I have assessed machine learning supervised classification and unsupervised clustering algorithms for 2D geological problems. These methods can, with appropriate data, be expanded to 3D geological models. For example, the geometric properties of SVM classification models, discussed in Section 8.1.1.2, offers the opportunity to generate generalised geological models from limited observational data using only spatial coordinates. Smirnoff *et al.* (2008) showed that SVM was able to construct plausible 3D geological models based on data representing 3D coordinate space. In their study, Smirnoff *et al.* (2008) employed a RBF SVM to categorise drill hole log data and surface geological maps. The use of SVM circumvented the need to manually generate 3D geological models via the reconstruction of individual geological layers using surfaces interpolated from observed lithology data, which were then combined into a single model. In contrast, RF is likely to be a better choice of algorithm when integrating multivariate 3D geoscience data such as gravity, TMI and Airborne Electro-Magnetics. The use of MLAs for 3D modelling situations will only be limited by the availability and coverage of 3D data.

In this thesis, I only investigate the use of MLAs for geoscience classification problems. In a broad sense, the production of maps of mineralisation prospectivity is essentially a regression problem, i.e. estimate the probability that a given location is prospective for the target mineralisation. The use of MLAs for 2D mineralisation prospectivity mapping using ANN is well documented (e.g., Singer & Kouda 1996; Skabar 2007; Oh & Lee 2010), whereas, examples using SVM are less common (e.g., Abedi *et al.* 2012). The methodologies developed in this thesis can be easily modified to generate numeric outputs for the purpose of mineral prospectivity mapping.

8.3.3. Big Data

The rate at which digital data is collected, including geoscience data, is rapidly increasing (Kraut & Wettergreen 2010; Bhatia *et al.* 2013). This large volume of geoscience data, coupled with advances in data storage and transmission technologies, are leading to a situation where new and novel approaches for the analysis and interpretation of so called "Big Data" (Hey *et al.* 2009) are required to enhance our scientific understanding of complex Earth systems (UNCOVER 2012; Sellars *et al.* 2013). The elements of Big Data for scientific research encompass (Lehning *et al.* 2009; Bhatia *et al.* 2013): (1) data

acquisition; (2) data management; (3) data analysis and modelling; and (4) communicating findings and interpretations.

The machine learning methodologies developed and documented in this thesis provide efficient and practical means of managing, analysing and interpreting geoscientific Big Data. For example, data management encompasses application specific techniques for the preparation, integration and extraction/selection of large and varied geoscience data such that inputs contain a manageable and representative set of information with which to conduct analyses. The machine learning data analysis methods described herein, offer geoscientists convenient methods with which to generate accurate and plausible geological maps for use in a wide variety of applications. Furthermore, I have demonstrated that meaningful interpretations of geoscience Big Data can be achieved by interrogating MLA decision structures and visualising the statistical and spatial distributions of MLA outputs.

Recent developments in Australian Big Data research are being supported and facilitated by the National eResearch Collaboration Tools and Resources (NeCTAR 2011). NeCTAR aims to provide researchers with access to state-of-the-art Information and Communication Technology (ICT) infrastructure for Big Data analysis and modelling. Earlier this year, NeCTAR collaborations with CSIRO, Geoscience Australia and the National Computational Infrastructure established a Virtual Geophysics Laboratory (VGL; http://vgl.auscope.org). The VGL offers geoscientists an online portal to a cloud-based repository of Australian geophysical data and a suite of software tools for geophysical inversion (NeCTAR 2011). The MLA workflows and R code (Appendix F) I have developed as part of my research have the potential to be integrated into the VGL. This will offer geoscientists access to machine learning geological inference tools that can be applied to national cloud-based data via well-posed methodologies and workflows.

CHAPTER 9 – CONCLUSIONS

In this thesis, machine learning algorithms have been employed to generate accurate geological maps efficiently with robust estimates of uncertainty from multivariate geospatial data and limited observed data. Machine learning algorithm generated geological maps, in conjunction with the decision structures induced by these algorithms, were used to formulate meaningful interpretations of complex geological phenomena.

The focus of my research was to conduct a robust assessment of the efficacy of machine learning algorithms to integrate geological, geophysical and geochemical data for supervised lithology/lithostratigraphy classification problems. The findings documented in this thesis provide new insights into the advantages and disadvantages associated with the use of different machine learning algorithms, representing five general machine learning strategies, for complex geological mapping applications. As a result, clear guidelines for the use of machine learning algorithms by non-experts for practical geoscience applications have been established.

I conclude that Random Forests is a good first-choice algorithm for predicting spatially varying geological phenomena, such as lithologies, from disparate geospatial data. In this thesis, I have identified a number of key characteristics of this particular algorithm that support this conclusion:

1a. Random Forests is insensitive to variations in its parameters and is thus straightforward to train unlike other popular machine learning algorithms such as Artificial Neural Networks and Support Vector Machines.

1b. Random Forests is very competitive with respect to other machine learning algorithms in terms of computational cost.

1c. Random Forests generates as accurate or significantly more accurate predictions of classes representing lithologic/lithostratigraphic units from geospatial data when compared to the other MLAs trialled in this thesis.

1d. The adaptive-nearest neighbour architecture of Random Forests allows it to generate efficiently and implicitly an ensemble of classifiers that represent localised (in variable space) models that exploit spatial-context.

2. Prediction uncertainty, derived from estimates of Random Forests class membership probabilities, is a reliable indication of ambiguous and/or erroneous classifications.
3. Random Forests provides in-built methods for interrogating classification model decision structures, such as variable importance and partial dependence, which can be used to gain an understanding of the interactions between geoscientific variables and the target classes under investigation as a means of facilitating the interpretation of complex geological phenomena.

In conjunction with identifying Random Forests as a good first-choice for the classification of spatially distributed geological phenomena, the research documented in this thesis describes in detail, three fundamental stages of the machine learning workflow (1) data pre-processing, (2) machine learning algorithm training and (3) prediction evaluation. I have identified key considerations for the execution of these stages when implementing MLAs for practical geoscience applications. These considerations acknowledge and address the challenges faced by geoscientists when employing machine learning algorithms for geospatial classification applications.

I have demonstrated that high Random Forests uncertainty provides an indication of geological transition zones such as faults and lithological contacts and regions. Random Forests classifications have been used to identify previously unmapped lithological units representing zones of intense hydrothermal alteration, in turn, providing opportunities to target ore deposits. In addition, the novel use of unsupervised clustering algorithms, specifically Self-Organising Maps, has been demonstrated in this thesis as means of providing robust methods for the preparation and analysis of geospatial data. Self-Organising Maps was shown to offer methods for significantly improving classification accuracy and interrogating geologically meaningful sub-classes within volcanic lithologies. These sub-classes were shown to represent compositional variations due to different phases of magmatism and contrasting zones of volcanic-hosted massive sulfide-related hydrothermal alteration.

I have developed R code to implement machine learning algorithms using the methodologies detailed in this thesis. This code can be freely distributed and/or modified for specific practical geoscience applications. The R code developed during the course of this thesis can be integrated into existing national research platforms such as the Virtual

Geophysics Laboratory, thus, providing the wider geoscience community with accessible and robust methods for the analysis and interpretation of large amounts of geoscience data as a means of addressing challenging geoscientific problems.

Recent technological advances in geoscience data capture and storage have significantly increased the volume and variety of digital information available to geoscientists. These data, coupled with the large amounts of pre-existing data, present challenges to large scale manual and/or deterministic approaches to geological mapping. I have demonstrated ways in which machine learning algorithms offer geoscientists novel tools for computer-assisted data integration and pattern recognition as a means of addressing these challenges. The methodology and outputs generated as a result of my research will be of significant practical benefit to a broad range of geoscience applications, such as: ore deposit targeting and modelling; reconstruction of tectonic histories; geohazard risk assessment; and hydrological modelling.

REFERENCES

Abedi, M., Norouzi, G.-H. & Bahroudi, A. 2012, 'Support Vector Machine for multi-classification of mineral prospectivity areas', *Computers & Geosciences*, vol. 46, pp. 272-283.

Aha, D.W. 1997, 'Lazy learning', *Artificial Intelligence Review*, vol. 11, no. 1-5, pp. 7-10.

Al-Aidaroos, K.M., Bakar, A.A. & Othman, Z. 2012, 'Medical data classification with Naive Bayes approach', *Information Technology Journal*, vol. 11, pp. 1166-1174.

Al-Shayea, Q.K. 2011, 'Artificial Neural Networks in medical diagnosis', *International Journal of Computer Science Issues*, vol. 8, no. 2, pp. 150-154.

Ali, S.S.M., Joshi, N., George, B. & Vanajakshi, L. 2012, 'Application of Random Forest algorithm to classify vehicles detected by a multiple inductive loop system', in *15th International IEEE Conference on Intelligent Transportation Systems*, September 2012, Anchorage, Alaska, pp. 491-495.

Allen, R.L., Weihed, P., Blundell, D., Crawford, T., Davidson, G., Galley, A., Gibson, H., Hannington, M., Herzig, P., Large, R., Lentz, D., Maslennikov, V.M., S., Peter, J. & Torno, F. 2002, 'Global comparisons of volcanic-associated massive sulphide districts', *Geological Society, London, Special Publications*, vol. 204, no. 1, pp. 13-37.

An, P. & Chung, C.-J.F. 1994, 'Neural network approach for geological mapping: technical background and case study', *Canadian Journal of Remote Sensing*, vol. 20, no. 3, pp. 293-301.

Anderson, M.J. & Ferree, C.E. 2010, 'Conserving the stage: climate change and the geophysical underpinnings of species diversity', *PLoS One*, vol. 5, no. 7, p. e11554.

Anselin, L. 1995, 'Local indicators of spatial association – LISA', *Geographical Analysis*, vol. 27, no. 2, pp. 93-115.

Aster, R.C., Borchers, B. & Thurber, C.H. 2005, *Parameter Estimation and Inverse Problems*, International Geophysical Series, Elsevier Academic Press, Burlington, Massachusetts, p. 376.

Åström, F. & Koker, R. 2011, 'A parallel neural network approach to prediction of Parkinson's Disease', *Expert Systems with Applications*, vol. 38, no. 10, pp. 12470-12474.

Atkeson, C.G., Moore, A.W. & Schaal, S. 1997, 'Locally weighted learning', *Artificial Intelligence Review*, vol. 11 no. 1-5, pp. 11-73.

Atkinson, P.M. & Lewis, P. 2000, 'Geostatistical classification for remote sensing: an introduction', *Computers & Geosciences*, vol. 26, no. 4, pp. 361-371.

Atkinson, P.M. & Tate, N.J. 2000, 'Spatial scale problems and geostatistical solutions: a review', *Professional Geographer*, vol. 52, no. 4, pp. 607-623.

Augustinus, P. & Colhoun, E.A. 1986, 'Glacial history of the upper Pieman and Boco valleys, western Tasmania', *Australian Journal of Earth Sciences*, vol. 33, no. 2, pp. 181-191.

Australian Geological Survey Organisation 1994, *Project 617, Broken Hill NSW 1994*, Geoscience Australia, Canberra, Australian Capital Territory.

Backer, E. & Jain, A.K. 1981, 'A clustering performance-measure based on fuzzy set decomposition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, no. 1, pp. 66-75.

Baddeley, A. & Turner, R. 2005, 'spatstat: an R package for analyzing spatial point patterns', *Journal of Statistical Software*, vol. 12, no. 6, pp. 1-42.

Baldwin, J.L., Otte, D.N., Wheatley, C.L. & Tomich, J.F. 1989, 'Computer emulation of human mental processes; application of neural network simulators to problems in well log interpretation', in *SPE Annual Technical Conference and Exhibition*, October 1989, San Antonio, Texas, vol. 64, pp. 481-493.

Banfield, R.E., Hall, L.O., Bowyer, K.W. & Kegelmeyer, W.P. 2007, 'A comparison of decision tree ensemble creation techniques', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 173-180.

Banks, M.R. & Baillie, P.W. 1989, 'Late Cambrian–Devonian', in C.F. Burrett & E.L. Martin (eds), *Geology and Mineral Resources of Tasmania*, Special Publication, Geological Society of Australia, Brisbane, Queensland, vol. 15, pp. 182-237.

Bárdossy, G. & Fodor, J. 2001, 'Traditional and new ways to handle uncertainty in geology', *Natural Resources Research*, vol. 10, no. 3, pp. 179-187.

Bedini, E. 2009, 'Mapping lithology of the Sarfartoq carbonatite complex, southern West Greenland, using HyMap imaging spectrometer data', *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1208-1219.

Bedini, E. 2012, 'Mapping alteration minerals at Malmbjerg molybdenum deposit, central East Greenland, by Kohonen Self-Organizing Maps and matched filter analysis of HyMap data', *International Journal of Remote Sensing*, vol. 33, no. 4, pp. 939-961.

Bellman, R.E. 1961, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, New Jersey, p. 255.

Bernatzki, A., Eppler, W. & Gemmeke, H. 1996, 'Interpretation of Neural Networks for Classification Tasks', in *European Congress on Intelligent Techniques and Soft Computing*, September 1996, Aachen, Germany, vol. 1, p. 5.

Berry, R. 1989, 'The history of movement on the Henty Fault Zone, western Tasmania: An analysis of fault striations', *Australian Journal of Earth Sciences*, vol. 36, pp. 189-205.

Berry, R.F. & Crawford, A.J. 1988, 'The tectonic significance of Cambrian allochthonous mafic ultramafic complexes in Tasmania', *Australian Journal of Earth Sciences*, vol. 35, no. 4, pp. 523-533.

Bezdek, J.C. 1981, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Advanced Applications in Pattern Recognition, Springer, New York, p. 256

Bhatia, A., Anuradha, B. & Gaurav, V. 2013, 'Big Data – a review', *International Journal of Engineering Sciences & Research Technology*, vol. 2, no. 8, pp. 2102-2106.

Bhatt, A. & Helle, H.B. 2002, 'Determination of facies from well logs using modular neural networks', *Petroleum Geoscience*, vol. 8, no. 3, pp. 217-228.

Bierlein, F.P., Fraser, S.J., Brown, W.M. & Lees, T. 2008, 'Advanced methodologies for the analysis of databases of mineral deposits and major faults', *Australian Journal of Earth Sciences*, vol. 55, no. 1, pp. 79-99.

Binbin, H., Ying, C., Cuihua, C., Jianhua, C. & Yue, L. 2011, 'Uncertainty mapping method for mineral resources prospectivity integrating multi-source geology spatial data sets and evidence reasoning model', in *Proceedings of the 19th International Conference on Geoinformatics*, June 2011, Shanghai, China, p. 5.

Bivand, R. & Rundel, C. 2012, *rgeos: interface to geometry engine - open source (GEOS), R package version 0.2-7*, <http://CRAN.R-project.org/package=rgeos>.

Bivand, R.S., Pebesma, E.J. & Gomez-Rubio, V. 2008, *Applied Spatial Data Analysis with R*, Springer, New York, p. 378.

Blanzieri, E. & Melgani, F. 2008, 'Nearest neighbor classification of remote sensing images with the maximal margin principle', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 6, pp. 1804-1811.

Blanzieri, E., Melgani, F. 2006, An Adaptive SVM nearest neighbor classifier for remotely sensed imagery, in *IEEE International Geoscience and Remote Sensing Symposium*, July-August 2006, Denver, Colorado, vols. 1-8, pp. 3931-3934

Boardman, J.W. 1989, 'Inversion of imaging spectrometry data using singular value decomposition', in *IEEE International Geoscience and Remote Sensing Symposium*, July 1989, Vancouver, Canada, vol. 4, pp. 2069-2072.

Bodin, T., Sambridge, M., Tkalcic, H., Arroucau, P., Gallagher, K. & Rawlinson, N. 2012, 'Transdimensional inversion of receiver functions and surface wave dispersion', *Journal of Geophysical Research-Solid Earth*, vol. 117, p. B02301.

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M. & Stum, A.K. 2008, 'Landsat Spectral Data for Digital Soil Mapping', in A.E. Hartemink, McBratney, A. B. and Mendonça-Santos, M. de L. (ed.), *Digital Soil Mapping with Limited Data*, Springer, Dordrecht, Netherlands, pp. 192-202.

Boisvert, J., Manchuk, J. & Deutsch, C. 2009, 'Kriging in the presence of locally varying anisotropy using non-Euclidean distances', *Mathematical Geosciences*, vol. 41, no. 5, pp. 585-601.

Boisvert, J.B. & Deutsch, C.V. 2011, 'Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances', *Computers & Geosciences*, vol. 37, no. 4, pp. 495-510.

Bottou, L. & Vapnik, V.N. 1992, 'Local learning algorithms', *Neural Computation*, vol. 4, pp. 888-900.

Bousquet, O., Boucheron, S. & Lugosi, G. 2004, 'Introduction to Statistical Learning Theory', in O. Bousquet, U. von Luxburg & G. Ratsch (eds), *Advanced Lectures on Machine Learning Lecture Notes in Artificial Intelligence 3176*, Springer, Heidelberg, Germany, pp. 169-207.

Brazdil, P.B. & Henery, R.J. 1994, 'Analysis of Results', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 175-212.

Breiman, L. 1996, 'Bagging predictors', *Machine Learning*, vol. 24, no. 2, pp. 123-140.

Breiman, L. 2001, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5-32.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. 1984, *Classification and Regression Trees*, Wadsworth & Brooks/Cole Statistics/Probability Series, Wadsworth International Group, Pacific Grove, California, p. 358.

Brion, G.M., Neelakantan, T.R. & Lingireddy, S. 2002, 'A neural-network-based classification scheme for sorting sources and ages of fecal contamination in water', *Water Research*, vol. 36, no. 15, pp. 3765-3774.

Brodaric, B., Gahegan, M. & Harrap, R. 2004, 'The art and science of mapping: computing geological categories from field data', *Computers & Geosciences*, vol. 30, no. 7, pp. 719-740.

Brown, A.V. 1986, *Geology of the Dundas-Mt Lindsay-Mt Youngbuck region*, GSB62, Tasmanian Geological Survey Bulletin, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/GSB62/GSB62.pdf>.

Brown, A.V. 1989, 'Eo-Cambrian–Cambrian', in C.F. Burrett & E.L. Martin (eds), *Geology and Mineral Resources of Tasmania*, Special Publication Geological Society of Australia, Brisbane, Queensland, vol. 15, pp. 47-83.

Brown, D.G. 1998, 'Classification and boundary vagueness in mapping presettlement forest types', *International Journal of Geographical Information Science*, vol. 12, no. 2, pp. 105-129.

Brown, G. & Mies, B.A. 2012, *Vegetation Ecology of Socotra*, Springer, Dordrecht, Netherlands, p. 379.

Buckley, P.M., Moriarty, T., Needham, J. & compilers 2002, *Broken Hill Geoscience Database*, 2nd edn, Geological Survey of New South Wales, Sydney. New South Wales.

Bue, B.D. & Stepinski, T.F. 2007, 'Machine detection of Martian impact craters from digital topography data', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 265-274.

Burges, C.J.C. 1998, 'A tutorial on Support Vector Machines for pattern recognition', *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167.

Burl, M.C., Asker, L., Smyth, P., Fayyad, U., Perona, P., Crumpler, L. & Aubele, J. 1998, 'Learning to recognize volcanoes on Venus', *Machine Learning*, vol. 30, no. 2, pp. 165-194.

Burrough, P.A. & McDonnell, R.A. 1998, *Principles of Geographical Information Systems*, 2nd edn, Oxford University Press, New York, p. 356.

Buselli, G. & Lu, K.L. 2001, 'Groundwater contamination monitoring with multichannel electrical and electromagnetic methods', *Journal of Applied Geophysics*, vol. 48, no. 1, pp. 11-23.

Camps-Valls, G., Gómez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martin-Guerrero, J. & Moreno, J. 2003, 'Support vector machines for crop classification using hyperspectral data', in *Pattern Recognition And Image Analysis, Lecture Notes in Computer Science 2652*, Springer, Heidelberg, Germany, pp. 134-141.

Carneiro, C.C., Fraser, S.J., Croacutesta, A.P., Silva, A.M. & Barros, C.E.M. 2012, 'Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon', *Geophysics*, vol. 77, no. 4, pp. K17-K24.

Carranza, E.J.M. 2011, 'From predictive mapping of mineral prospectivity to quantitative estimation of number of undiscovered prospects', *Resource Geology*, vol. 61, no. 1, pp. 30-51.

Caruana, R. & Mizil, A. 2006, 'An empirical comparison of supervised learning algorithms', in *Proceedings of the 23rd International Conference On Machine Learning*, June 2006, Pittsburgh, Pennsylvania, pp. 161-168.

Ceamanos, X., Waske, B., Benediktsson, J., Chanussot, J. & Sveinsson, J. 2009, 'Ensemble strategies for classifying hyperspectral remote sensing data', in *8th International Workshop on Multiple Classifier Systems*, June 2009, Reykjavik, Iceland, pp. 62-71.

Chakraborty, B., Mahale, V., de Sousa, C. & Das, P. 2004, 'Seafloor classification using echo-waveforms: a method employing hybrid neural network architecture', *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 3, pp. 196-200.

Chipman, H.A., George, E.I. & McCulloch, R.E. 2010, 'BART: Bayesian additive regression trees', *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266-298.

Cocks, T., Jenssen, R., Stewart, A., Wilson, I. & Shields, T. 1998, 'The HyMap™ airborne hyperspectral sensor: the system, calibration and performance', in M. Schaepman, D. Schläpfer & K.I. Itten (eds), *1st European Association of Remote Sensing*

Laboratories Workshop on Imaging Spectrometry, October 1998, Zurich, Switzerland, pp. 37-42.

Cohen, J. 1960, 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37-46.

Congalton, R.G. & Green, K. 1998, *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, CRC Press, Boca Raton, Florida, p. 160.

Corbett, K.D. & Komyshan, P. 1989, *Geology of the Hellyer–Mt Charter area*, Mt Read Volcanics Project, Geological Report 1, Tasmanian Department of Mines, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/MRV1/MRV1.pdf>.

Corbett, K.D. & Solomon, M. 1989, 'Cambrian Mt Read Volcanics and Associated Mineral Deposits', in C.F. Burrett & E.L. Martin (eds), *Geology and Mineral Resources of Tasmania*, Special Publication Geological Society of Australia, Brisbane, Queensland, vol. 15, pp. 84-153.

Cortes, C. & Vapnik, V. 1995, 'Support-vector networks', *Machine Learning*, vol. 20, no. 3, pp. 273-297.

Cover, T. & Hart, P. 1967, 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27.

Cracknell, M.J., Reading, A. & McNeill, A. in press, 2014, 'Mapping geology and volcanic-hosted massive sulphide alteration in the Hellyer–Mt charter region, Tasmania, using Random Forests™ and Self-Organising Maps', *Australian Journal of Earth Sciences*.

Cracknell, M.J. & Reading, A.M. 2013, 'The upside of uncertainty: identification of lithology contact zones from airborne geophysics and satellite data using Random Forests and Support Vector Machines', *Geophysics*, vol. 78, no. 3, pp. WB113–WB126.

Cracknell, M.J. & Reading, A.M. 2014, 'Geological mapping using remote sensing data: a comparison of five machine learning algorithms, their response to variations in the

spatial distribution of training data and the use of explicit spatial information', *Computers & Geosciences*, vol. 63, pp. 22-33.

Crawford, A.J. & Berry, R.F. 1992, 'Tectonic implications of Late Proterozoic–Early Palaeozoic igneous rock associations in western Tasmania', *Tectonophysics*, vol. 214, no. 1-4, pp. 37-56.

Crawford, A.J., Corbett, K.D. & Everard, J.L. 1992, 'Geochemistry of the Cambrian volcanic-hosted massive sulfide-rich Mt Read Volcanics, Tasmania, and some tectonic implications', *Economic Geology and the Bulletin of the Society of Economic Geologists*, vol. 87, no. 3, pp. 597-619.

Cudahy, T., Caccetta, M. & Thomas, M. 2012, *National ASTER Map of Australia*, 1 edn, Geoscience Australia, Canberra, Australian Capital Territory.

Cutler, A. & Stevens, J.R. 2006, 'Random forests for microarrays', *Methods in Enzymology*, vol. 411, pp. 422-432.

Cutler, D.R., Edwards, T.C., Jr., Beard, K.H., Cutler, A. & Hess, K.T. 2007, 'Random forests for classification in ecology', *Ecology*, vol. 88, no. 11, pp. 2783-2792.

Dai, H. 2003, 'Application of multilayer perceptrons to earthquake seismic analysis', in W. Sandham & M. Leggett (eds), *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 287-304.

Danese, M., Lazzari, M. & Murgante, B. 2008, 'Kernel density estimation methods for a geostatistical approach in seismic risk analysis: the case study of Potenza hilltop town (Southern Italy)', in *Proceeding of the International Conference on Computational Science and its Applications*, June-July 2008, Perugia, Italy, pp. 415-429.

Das, P. & Iyer, S. 2009, 'Geochemical characterization of oceanic basalts using Artificial Neural Network', *Geochemical Transactions*, vol. 10, no. 1, pp. 1-11.

Dauth, C. 1997, 'Airborne magnetic, radiometric and satellite imagery for regolith mapping in the Yilgarn Craton of Western Australia', *Exploration Geophysics*, vol. 28, no. 2, pp. 199-203.

De'ath, G. 2007, 'Boosted trees for ecological modeling and prediction', *Ecology*, vol. 88, no. 1, pp. 243-251.

De, C. & Chakraborty, B. 2009, 'Acoustic characterization of seafloor sediment employing a hybrid method of neural network architecture and fuzzy algorithm', *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 743-747.

Dêbska, B. & Guzowska-Œwider, B. 2011, 'Application of artificial neural network in food classification', *Analytica Chimica Acta*, vol. 705, no. 1-2, pp. 283-291.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S. & McLoone, S. 2013, 'Principal component analysis on spatial data: an overview', *Annals of the Association of American Geographers*, vol. 103, no. 1, pp. 106-128.

Denison, D.G.T., Holmes, C.C., Mallick, B.K. & Smith, A.F.M. 2002, *Bayesian Methods for Nonlinear Classification and Regression*, Wiley Series in Probability and Statistics, Wiley, Chichester, West Sussex, p. 296.

Depeursinge, A., Iavindrasana, J., Hidki, A., Cohen, G., Geissbuhler, A., Platon, A., Poletti, P.-A. & Müller, H. 2010, 'Comparative performance analysis of state-of-the-art classification algorithms applied to lung tissue categorization', *Journal of Digital Imaging*, vol. 23, no. 1, pp. 18-30.

Díaz-Uriarte, R. & De Andres, S.A. 2006, 'Gene selection and classification of microarray data using random forest', *BMC Bioinformatics*, vol. 7, no. 1, p. 3.

Dietterich, T. & Wettschereck, D. 1994, 'Locally adaptive nearest neighbor algorithms', *Advances in Neural Information Processing Systems*, vol. 6, pp. 184-191.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. 2011, *e1071: misc. functions of the department of statistics (e1071), R package version 1.5-26*, <http://CRAN.R-project.org/package=e1071>.

Ding, C.H. & Dubchak, I. 2001, 'Multi-class protein fold recognition using support vector machines and neural networks', *Bioinformatics*, vol. 17, no. 4, pp. 349-358.

Ding, W., Jiamthapthaksin, R., Parmar, R., Jiang, D., Stepinski, T.F. & Eick, C.F. 2008, 'Towards region discovery in spatial datasets', in *Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, May, 2008, Osaka, Japan, pp. 88-99.

Dohm, J.M., Wang, R., Dalton, J.B., Scharenbroich, L., Hare, T.M., Castano, R. & Baker, V.R. 2007, 'Are the rock compositions of the ancient mountain range of Mars, Thaumasia highlands, distinct from Tharsis lavas?: machine learning evaluation of TES data and implications on early evolution of Mars', in *NASA Science Technology Conference*, June 2007, University College, Maryland, pp. 1-6.

Domingos, P. & Pazzani, M. 1997, 'On the optimality of the simple Bayesian classifier under zero-one loss', *Machine Learning*, vol. 29, no. 2-3, pp. 103-130.

Draskovits, P. & Laszlo, V. 2005, 'Indication of groundwater contamination with the induced-polarization (IP) method', in D.K. Butler (ed.), *Near-Surface Geophysics: Investigations in Geophysics*, vol. 13, pp. 551–562.

Dubois, M.K., Bohling, G.C. & Chakrabarti, S. 2007, 'Comparison of four approaches to a rock facies classification problem', *Computers & Geosciences*, vol. 33, no. 5, pp. 599-617.

Duda, R.O. & Hart, P.E. 1973, *Pattern classification and scene analysis*, Wiley, Chichester, West Sussex, p. 512.

Dumais, S., Platt, J., Heckeman, D. & Sahami, M. 1998, 'Inductive learning algorithms and representations for text categorization', in *Proceedings of the 3rd International Conference on Information and Knowledge Management*, November 1998, Bethesda, Maryland, pp. 148-155.

Durning, W.P., Polis, S.R., Frost, E.G. & Kaiser, J.V. 1998, *Integrated Use of Remote Sensing and GIS for Mineral Exploration - Final Report*, ARC-SDSU-002-97, Affiliated Research Center, San Diego State University, San Diego, California, <http://www.gis.usu.edu/docs/data/nasa_arc/nasa_arc97/SDSU/LaCuesta.pdf>.

Duro, D.C., Franklin, S.E. & Dubé, M.G. 2012, 'A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery', *Remote Sensing of Environment*, vol. 118, pp. 259-272.

Efron, B. 1983, 'Estimating the error rate of a prediction rule - improvement on cross-validation', *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316-331.

Ehret, B. 2010, 'Pattern recognition of geophysical data', *Geoderma*, vol. 160, no. 1, pp. 111-125.

El-Naqa, I., Yang, Y., Wernick, M.N., Galatsanos, N.P. & Nishikawa, R.M. 2002, 'A support vector machine approach for detection of microcalcifications', *IEEE Transactions on Medical Imaging*, vol. 21, no. 12, pp. 1552-1563.

El-Sebakhy, E.A., Asparouhov, O., Abdulraheem, A., Wu, D., Latinski, K. & Spries, W. 2010, 'Data mining in identifying carbonate litho-facies from well logs based from extreme learning and support vector machines', in *Proceedings of the 9th Middle East Geoscience Conference & Exhibition*, March 2008, Manama, Bahrain, pp. 1-17.

Emery, X. & Gonzalez, K.E. 2007, 'Incorporating the uncertainty in geological boundaries into mineral resources evaluation', *Journal of the Geological Society of India*, vol. 69, no. 1, pp. 29-38.

Essenreiter, R., Karrenbach, M. & Treitel, S. 2003, 'Identification and Suppression of Multiple Reflections in Marine Seismic Data with Neural Networks', in W. Sandham & M. Leggett (eds), *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 71-88.

Farhad, S.G., Peyman, M. & Parvin, H. 2012, 'Application of decision tree algorithm for data mining in healthcare operations: a case study', *International Journal of Computer Applications*, vol. 52, no. 6, pp. 21-26.

Feng, C. & Michie, D. 1994, 'Machine Learning Rules and Trees', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 50-83.

Feyyad, U.M. 1996, 'Data mining and knowledge discovery: making sense out of data', *IEEE Expert*, vol. 11, no. 5, pp. 20-25.

Fisher, P.F. 1999, 'Models of uncertainty in spatial data', *Geographical Information Systems*, vol. 1: pp. 191-205.

Fix, E. & Hodges, J.L. 1951, *Discriminatory analysis. Nonparametric discrimination; Consistency properties*, Technical Report 4, U.S. Air Force, School of Aviation Medicine, Randolph Field, Texas.

Foody, G.M. 2009, 'Sample size determination for image classification accuracy assessment and comparison', *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273-5291.

Foody, G.M. & Mathur, A. 2004, 'A relative evaluation of multiclass image classification by support vector machines', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335-1343.

Fotheringham, A.S. 2009, '"The problem of spatial autocorrelation" and local spatial statistics', *Geographical Analysis*, vol. 41, no. 4, pp. 398-403.

Fotheringham, A.S., Brunsdon, C. & Charlton, M. 2002, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley, Chichester, West Sussex, p. 284.

Frank, A. & Asuncion, A. 2010, *UCI Machine Learning Repository*, University of California, Department of Information and Computer Science, <http://archive.ics.uci.edu/ml>.

Franklin, J.M., H.L. Gibson, Jonasson, I.R. & Galley, A.G. 2005, 'Volcanogenic Massive Sulphide Deposits', in J.W. Hedenquist, J.F.H. Thompson, R.J. Goldfarb & J.P. Richards (eds), *Economic Geology, 100th Anniversary Volume*, Society Of Economic Geologists, Inc., Littleton, Colorado, pp. 523 - 560.

Franklin, S.E., Wulder, M.A. & Lavigne, M.B. 1996, 'Automated derivation of geographic window sizes for use in remote sensing digital image texture analysis', *Computers & Geosciences*, vol. 22, no. 6, pp. 665-673.

Fraser, S.J. & Dickson, B.L. 2007, 'A new method for data integration and integrated data interpretation: Self-Organising Maps', in *Proceedings of the 5th Decennial International Conference on Mineral Exploration*, September 2007, Toronto, Canada, pp. 907-910.

Freund, Y. & Schapire, R. 1996, 'Experiments with a new boosting algorithm', in *Proceedings of the Thirteenth International Conference on Machine Learning*, July 1996, Bari, Italy, pp. 148-156.

Freund, Y. & Schapire, R.E. 1997, 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139.

Friedman, J.H. 1994, *Flexible Metric Nearest Neighbor Classification* 113, Department of Statistics, Stanford University, Stanford, California, <http://statistics.stanford.edu/~ckirby/techreports/LCS/LCS%20113.pdf>.

Friedman, J.H. 1997, 'On bias, variance, 0/1—loss, and the Curse-of-Dimensionality', *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55-77.

Friedman, J.H. 2001, 'Greedy function approximation: a gradient boosting machine', *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232.

Friedman, N., Geiger, D. & Goldszmidt, M. 1997, 'Bayesian network classifiers', *Machine Learning*, vol. 29, no. 2-3, pp. 131-163.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. & Haussler, D. 2000, 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics*, vol. 16, no. 10, pp. 906-914.

Gabriel, D. 2007, 'Mapping Petrological Patterns in the Regouffe Granite by Integrating Geochemical, VNIR-SWIR and Gamma-Ray Spectrometry Data', Master of

Science, International Institute for Geo-information Science and Earth Observation, Ensched, The Netherlands, p. 85, <http://www.itc.nl/library/papers_2007/msc/aes/data.pdf>.

Gahegan, M. 2000, 'On the application of inductive machine learning tools to geographical analysis', *Geographical Analysis*, vol. 32, no. 2, pp. 113-139.

Galley, A., Hannington, M. & Jonasson, I. 2007, 'Volcanogenic massive sulphide deposits', *Mineral Deposits of Canada: A Synthesis of Major Deposit-Types, District Metallogeny, the Evolution of Geological Provinces, and Exploration Methods.* no. 5, pp. 141-161.

Galley, A.G. 1995, 'Targeting vectoring using lithogeochemistry: Applications to the exploration for volcanic-hosted massive sulphide deposits', *Canadian Institute of Mining Bulletin*, vol. 88, no. 990, pp. 15–27.

Gelfort, R. 2006, 'On Classification of Logging Data', Doctor of Philosophy, Clausthal University of Technology, Clausthal, Germany, p. 131.

Geman, S., Bienenstock, E. & Doursat, R. 1992, 'Neural networks and the bias variance dilemma', *Neural Computation*, vol. 4, no. 1, pp. 1-58.

Gemmell, J.B. & Fulton, R. 2001, 'Geology, genesis, and exploration implications of the foot wall and hanging-wall alteration associated with the Hellyer volcanic-hosted massive sulfide deposit, Tasmania, Australia', *Economic Geology and the Bulletin of the Society of Economic Geologists*, vol. 96, no. 5, pp. 1003-1035.

Geoscience Australia 2010, *Geomagnetism - AGRF 2010 model field values calculations*, viewed May, 2012, <http://www.ga.gov.au/oracle/geomag/agrfform.jsp>.

Getis, A. 2010, 'Spatial Autocorrelation', in M.M. Fisher & A. Getis (eds), *Handbook of Applied Spatial Analysis: Software, Tools, Methods and Applications*, Springer-Verlag, Berlin, Germany, pp. 255-278.

Getis, A. & Ord, J.K. 1992, 'The analysis of spatial association by use of distance statistics', *Geographical Analysis*, vol. 24, no. 3, pp. 189-206.

Ghimire, B., Rogan, J., Galiano, V.R., Panday, P. & Neeti, N. 2012, 'An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA', *GIScience & Remote Sensing*, vol. 49, no. 5, pp. 623-643.

Ghimire, B., Rogan, J. & Miller, J. 2010, 'Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic', *Remote Sensing Letters*, vol. 1, no. 1, pp. 45-54.

Ghosh, S., Stepinski, T.F. & Vilalta, R. 2010, 'Automatic annotation of planetary surfaces with geomorphic labels', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 1, pp. 175-185.

Gibson, G.M., Rubenach, M.J., Neumann, N.L., Southgate, P.N. & Hutton, L.J. 2008, 'Syn- and post-extensional tectonic activity in the Palaeoproterozoic sequences of Broken Hill and Mt Isa and its bearing on reconstructions of Rodinia', *Precambrian Research*, vol. 166, no. 1-4, pp. 350-369.

Gifford, C.M. & Agah, A. 2010, 'Collaborative multi-agent rock facies classification from wireline well log data', *Engineering Applications of Artificial Intelligence*, vol. 23, no. 7, pp. 1158-1172.

Gifford, C.M. & Agah, A. 2012, 'Subglacial water presence classification from polar radar data', *Engineering Applications of Artificial Intelligence*, vol. 25, no. 4, pp. 853-868.

Glasziou, P. & Hilden, J. 1989, 'Test selection measures', *Medical Decision Making*, vol. 9, no. 2, pp. 133-141.

Gonçalves, M.L., Netto, M.L.A., Costa, J.A.F. & Zullo Jr, J. 2008, 'An unsupervised method of classifying remotely sensed images using Kohonen Self-Organizing Maps and agglomerative hierarchical clustering methods', *Journal of Remote Sensing*, vol. 29, pp. 3171–3207.

Gonzalez, R.C. & Woods, R.E. 2008, *Digital Image Processing*, 3rd edn, Prentice Hall Inc., Upper Saddle River, New Jersey, p. 954.

Goodchild, M.F., Chih-Chang, L. & Leung, Y. 1994, 'Visualizing Fuzzy Maps', in H.M. Hearnshaw & D.J. Unwin (eds), *Visualization in GIS*, Wiley, New York, pp. 158-167.

Goodchild, M.F. & Quattrochi, D.A. 1997, 'Scale, Multiscaling, Remote Sensing and GIS', in D.A. Quattrochi & M.F. Goodchild (eds), *Scale in Remote Sensing and GIS*, Lewis Publishers, Boca Raton, Florida, pp. 1-12.

Goodchild, M.F., Sun, G.Q. & Shiren, Y. 1992, 'Development and test of an error model for categorical data', *International Journal of Geographical Information Systems*, vol. 6, no. 2, pp. 87-104.

Gotway, C.A. & Young, L.J. 2002, 'Combining incompatible spatial data', *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 632-648.

Govindarajan, M. & Chandrasekaran, R.M. 2010, 'Evaluation of *k*-Nearest Neighbor classifier performance for direct marketing', *Expert Systems with Applications*, vol. 37, no. 1, pp. 253-258.

Granath, G. 1988, 'Pattern recognition in geochemical hydrocarbon exploration: a fuzzy approach', *Mathematical Geology*, vol. 20, no. 6, pp. 673-691.

Grebby, S., Naden, J., Cunningham, D. & Tansey, K. 2011, 'Integrating airborne multispectral imagery and airborne LiDAR data for enhanced lithological mapping in vegetated terrain', *Remote Sensing of Environment*, vol. 115, no. 1, pp. 214-226.

Groves, D.I., Murchie, H., Martin, E.L. & Wellington, H.K. 1972, *A century of tin mining at Mt Bischoff, 1871-1971*, Tasmanian Department of Mines, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/GSB54/GSB54.pdf>.

Gubbins, D. 2004, *Time Series Analysis and Inverse Theory for Geophysicists*, Cambridge University Press, Cambridge, p. 255.

Guisan, A. & Zimmermann, N.E. 2000, 'Predictive habitat distribution models in ecology', *Ecological Modelling*, vol. 135, no. 2–3, pp. 147-186.

Guyon, I. 2008, 'Practical feature selection: from correlation to causality', in F. Fogelman-Soulié, D. Perrotta, J. Piskorski & R. Steinberger (eds), *Mining Massive Data Sets for Security - Advances in Data Mining, Search, Social Networks and Text Mining, and their Applications to Security*, IOS Press, Amsterdam, Netherlands, vol. 19, pp. 27-43, <http://eprints.pascal-network.org/archive/00004038/01/PracticalFS.pdf>.

Guyon, I. 2009, 'A practical guide to model selection', in J. Marie (ed.), in *Proceedings of the Machine Learning Summer School*, Springer, Canberra, Australia, January 26 - February 6, p. 37, <http://eprints.pascal-network.org/archive/00005768/01/guyon-mlss.pdf>.

Hallberg, J.A. 1984, 'A geochemical aid to igneous rock type identification in deeply weathered terrain', *Journal of Geochemical Exploration*, vol. 20, no. 1, pp. 1-8.

Ham, J., Yangchi, C., Crawford, M.M. & Ghosh, J. 2005, 'Investigation of the Random Forest framework for classification of hyperspectral data', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492-501.

Hammer, B. & Villmann, T. 2007, 'How to process uncertainty in machine learning?', in M. Verleysen (ed.), *15th European Symposium on Artificial Neural Networks*, April 2007, Bruges, Belgium, pp. 79 - 90.

Haralick, R.M., Shanmugam, K. & Dinstein, I.H. 1973, 'Textural features for image classification', *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621.

Hart, D.I. 2003, 'Automated Picking of Seismic First-Arrivals with Neural Networks', in W. Sandham & M. Leggett (eds), *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 13-30.

Hastie, T. & Tibshirani, R. 1996, 'Discriminant adaptive nearest neighbor classification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 607-616.

Hastie, T., Tibshirani, R. & Friedman, J.H. 2009, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edn, Springer Series in Statistics, Springer, New York, p. 533.

Hechenbichler, K. & Schliep, K.P. 2004, 'Weighted k-Nearest-Neighbor Techniques and Ordinal Classification', vol. Discussion Paper 399, no. SFB 386, Ludwig-Maximilians University, Munich, Germany, <http://epub.ub.uni-muenchen.de/1769/1/paper_399.pdf>.

Heermann, P.D. & Khazenie, N. 1992, 'Classification of multispectral remote sensing data using a back-propagation neural network', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 30, no. 1, pp. 81-88.

Henery, R.J. 1994a, 'Classification', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 6-16.

Henery, R.J. 1994b, 'Methods for Comparison', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 107-124.

Hengl, T. 2009, *A Practical Guide to Geostatistical Mapping*, 2nd edn, Scientific and Technical Research Series Report, Office for Official Publications of the European Communities, Luxembourg, p. 270.

Hey, T., Tansley, S. & Tolle, K. (eds) 2009, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Washington, p. 252.

Hijmans, R.J. & van Etten, J. 2012, *raster: geographic analysis and modeling with raster data, R package version 2.0-12*, <http://CRAN.R-project.org/package=raster>.

Holden, E.-J., Dentith, M. & Kovesi, P. 2008, 'Towards the automated analysis of regional aeromagnetic data to identify regions prospective for gold deposits', *Computers & Geosciences*, vol. 34, no. 11, pp. 1505-1513.

Howell, E., Merenyi, E. & Lebofsky, L. 1994, 'Classification of asteroid spectra using a neural network', *Journal of Geophysical Research*, vol. 99, no. E5, pp. 10847-10865.

Hsu, C.-W., Chang, C.-C. & Lin, C.-J. 2010, 'A Practical Guide to Support Vector Classification', Department of Computer Science, National Taiwan University, Taipei, Taiwan, p. 16, <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.

Hsu, C.-W. & Lin, C.-J. 2002, 'A comparison of methods for multiclass support vector machines', *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415-425.

Hua, S. & Sun, Z. 2001, 'Support vector machine approach for protein subcellular localization prediction', *Bioinformatics*, vol. 17, no. 8, pp. 721-728.

Huang, C., Davis, L.S. & Townshend, J.R.G. 2002, 'An assessment of support vector machines for land cover classification', *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725-749.

Huang, H.P. & Fraser, D.C. 2000, 'Airborne resistivity and susceptibility mapping in magnetically polarizable areas', *Geophysics*, vol. 65, no. 2, pp. 502-511.

Huang, Y., Kangas, L.J. & Rasco, B.A. 2007, 'Applications of Artificial Neural Networks (ANNs) in Food Science', *Critical Reviews in Food Science and Nutrition*, vol. 47, no. 2, pp. 113-126.

Hughes, G.F. 1968, 'On the mean accuracy of statistical pattern recognizers', *IEEE Transactions on Information Theory*, vol. 14, pp. 55 - 63.

Illa, J., Alonso, J. & Marré, M. 2004, 'Nearest-Neighbours for time series', *Applied Intelligence*, vol. 20, no. 1, pp. 21-35.

Inzana, J., Kusky, T., Higgs, G. & Tucker, R. 2003, 'Supervised classifications of Landsat TM band ratio images and Landsat TM band ratio image with radar for geological interpretations of central Madagascar', *Journal of African Earth Sciences*, vol. 37, no. 1-2, pp. 59-72.

Jackson, J. 2005, 'The use of sub-audio magnetics (SAM) in gold exploration - examples from the Yilgarn Craton, WA', *Exploration Geophysics*, vol. 36, no. 2, pp. 163-169.

James, G.M. 2003, 'Variance and bias for general loss functions', *Machine Learning*, vol. 51, no. 2, pp. 115-135.

Japkowicz, N. & Stephen, S. 2002, 'The class imbalance problem: a systematic study', *Intelligent Data Analysis*, vol. 6, no. 7, pp. 429-450.

Joachims, T. 2002, *Learning to Classify Text Using Support Vector Machines - Methods, Theory, and Algorithms*, Kluwer Academic Publishers, Norwell, Massachusetts, p. 205.

John, G.H. & Langley, P. 1995, 'Estimating continuous distributions in Bayesian classifiers', in P. Besnard & S. Hanks (eds), in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, August 1995, Montreal, Quebec, Canada, pp. 338-345.

Jolliffe, I.T. 2002, *Principal Component Analysis*, 2nd edn, Springer Series in Statistics, Springer-Verlag, New York, p. 489.

Joshi, A.J., Porikli, F. & Papanikolopoulos, N. 2009, 'Multi-class active learning for image classification', in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, Miami, Florida, pp. 2372-2379.

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V. & Canu, S. 2004, 'Environmental data mining and modeling based on machine learning algorithms and geostatistics', *Environmental Modelling & Software*, vol. 19, no. 9, pp. 845-855.

Kanevski, M., Pozdnoukhov, A. & Timonin, V. 2009, *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*, CRC Press, Boca Raton, Florida, p. 368.

Karatzoglou, A., Meyer, D. & Hornik, K. 2006, 'Support Vector Machines in R', *Journal of Statistical Software*, vol. 15, no. 9, p. 28.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. 2004, 'kernlab - an S4 package for kernel methods in R', *Journal of Statistical Software*, vol. 11, no. 9, pp. 1-20.

Kaur, S. & Josan, G.S. 2011, 'Gurmukhi text extraction from image using Support Vector Machine (SVM)', *International Journal of Engineering Science and Technology*, vol. 3, no. 4, pp. 2977-2911.

Keitt, T.H., Bivand, R., Pebesma, E.J. & Rowlingson, B. 2012, *rgdal: bindings for the geospatial data abstraction library, R package version 0.7-8*, <http://CRAN.R-project.org/package=rgdal>.

Khan, I.Y., Zope, P.H. & Suralkar, S.R. 2013, 'Importance of Artificial Neural Network in Medical Diagnosis disease like acute nephritis disease and heart disease', *International Journal of Engineering Science and Innovative Technology*, vol. 2, no. 2, pp. 210-217.

Kiviluoto, K. 1996, 'Topology preservation in Self-Organizing Maps', in *IEEE International Conference on Neural Networks*, June 1996, Washington, DC, vol. 1, pp. 294-299.

Kohavi, R. 1995, 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, August 1995, Montreal, Quebec, vol. 2, pp. 1137-1143.

Kohavi, R. & Wolpert, D.H. 1996, 'Bias Plus variance decomposition for zero-one loss functions', in L. Saitta (ed.), in *13th International Machine Learning Conference*, July 1996, Bari, Italy, pp. 275-283.

Kohonen, T. 1982, 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69.

Kohonen, T. 2001, *Self-Organizing Maps*, Springer Series in Information Sciences, Springer-Verlag, Berlin, Germany, vol. 30, p. 501.

Koike, K., Matsuda, S., Suzuki, T. & Ohmi, M. 2002, 'Neural Network-based estimation of principal metal contents in the Hokuroku District, Northern Japan, for exploring Kuroko-type deposits', *Natural Resources Research*, vol. 11, no. 2, pp. 135-156.

Kotsiantis, S.B. 2007, 'Supervised machine learning: a review of classification techniques', *Informatica*, vol. 31, pp. 249-268.

Kovacevic, M., Bajat, B. & Gajic, B. 2010, 'Soil type classification and estimation of soil properties using support vector machines', *Geoderma*, vol. 154, no. 3-4, pp. 340-347.

Kovacevic, M., Bajat, B., Trivic, B. & Pavlovic, R. 2009, 'Geological units classification of multispectral images by using Support Vector Machines', in *IEEE International Conference on Intelligent Networking and Collaborative Systems*, November 2009, Barcelona, Spain, pp. 267-272 .

Kraut, A. & Wettergreen, D. 2010, 'Classification of Mars terrain using multiple data sources', in *NASA Conference on Intelligent Data Understanding*, October 2010, Mountain View, California, pp. 54-68.

Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J. & Goetz, A.F.H. 1993, 'The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data', *Remote Sensing of Environment*, vol. 44, no. 2–3, pp. 145-163.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C. & Engelhardt, A. 2012, *caret: Classification and Regression Training, R package version 5.15-023*, <http://CRAN.R-project.org/package=caret>.

Kumar, K. & Thakur, G.S.M. 2012, 'Advanced Applications of neural networks and artificial intelligence: a review', *International Journal of Information Technology and Computer Science*, vol. 4, no. 6, pp. 57-68.

Kuncheva, L. 2004, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, Hoboken, New Jersey, p. 376.

Kusky, T.M. & Ramadan, T.M. 2002, 'Structural controls on Neoproterozoic mineralization in the South Eastern Desert, Egypt: an integrated field, Landsat TM, and SIR-C/X SAR approach', *Journal of African Earth Sciences*, vol. 35, no. 1, pp. 107-121.

Lacassie, J.P., del Solar, J.R., Roser, B. & Hervé, F. 2006, 'Visualization of volcanic rock geochemical data and classification with artificial neural networks', *Mathematical Geology*, vol. 38, no. 6, pp. 697-710.

Landgrebe, T.C.W. & Paclik, P. 2010, 'The ROC skeleton for multiclass ROC estimation', *Pattern Recognition Letters*, vol. 31, no. 9, pp. 949-958.

Landis, J.R. & Koch, G.G. 1977, 'The Measurement of observer agreement for categorical data', *Biometrics*, vol. 33, no. 1, pp. 159-174.

Langer, H., Falsaperla, S., Masotti, M., Campanini, R., Spampinato, S. & Messina, A. 2009, 'Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy', *Geophysical Journal International*, vol. 178, no. 2, pp. 1132-1144.

Large, R.R., McPhie, J., Gemmell, J.B., Herrmann, W. & Davidson, G.J. 2001, 'The spectrum of ore deposit types, volcanic environments, alteration halos, and related exploration vectors in submarine volcanic successions: some examples from Australia', *Economic Geology*, vol. 96, no. 5, pp. 913-938.

Leaman, D.E. & Richardson, R.G. 1989, *The granites of west and north-west Tasmania-a geophysical interpretation*, Department of Mines, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/GSB66/GSB66.pdf>.

Lee, Y.-H., Wei, C.-P., Cheng, T.-H. & Yang, C.-T. 2012, 'Nearest-neighbor-based approach to time-series classification', *Decision Support Systems*, vol. 53, no. 1, pp. 207-217.

Lehning, M., Dawes, N., Bavay, M., Parlange, M., Nath, S. & Zhao, F. 2009, 'Instrumenting the Earth: Next-Generation Sensor Networks and Environmental Science', in T. Hey, S. Tansley & K. Tolle (eds), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Washington, pp. 45-51.

Lepistö, L., Kunttu, I. & Visa, A. 2006, 'Classification of natural rock images using classifier combinations', *Optical Engineering*, vol. 45, no. 9, pp. 097201-097201.

Leverington, D.W. 2010, 'Discrimination of sedimentary lithologies using Hyperion and Landsat Thematic Mapper data: A case study at Melville Island, Canadian High Arctic', *International Journal of Remote Sensing*, vol. 31, no. 1, pp. 233-260.

Leverington, D.W. & Moon, W.M. 2012, 'Landsat-TM-based discrimination of lithological units associated with the Purtuniq Ophiolite, Quebec, Canada', *Remote Sensing*, vol. 4, no. 5, pp. 1208-1231.

Lewin-Koh, J.N., Bivand, R., Pebesma, E.J., Archer, E., Baddeley, A., Bibiko, H.-J., Dray, S., Forrest, D., Friendly, M., Giraudoux, P., Golicher, D., Hausmann, V.G.R.P., Hufthammer, K.O., Jagger, T., Luque, S.P., MacQueen, D., Niccolai, A., Short, T., Stabler, B. & Turner, R. 2012, *maptools: tools for reading and handling spatial objects, R package version 0.8-16*, <http://CRAN.R-project.org/package=maptools>.

Li, C.-H., Kuo, B.-C., Lin, C.-T. & Huang, C.-S. 2012, 'A spatial-contextual Support Vector Machine for remotely sensed image classification', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 3, pp. 784-799.

Li, J. & Narayanan, R.M. 2004, 'Integrated spectral and spatial information mining in remote sensing imagery', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 673-685.

Li, S., Kwok, J.T., Zhu, H. & Wang, Y. 2003, 'Texture classification using the support vector machines', *Pattern Recognition*, vol. 36, no. 12, pp. 2883-2893.

Liaw, A. & Wiener, M. 2002, 'Classification and Regression by randomForest', *RNews*, vol. 2, no. 3, pp. 18-22.

Lim, C.P., Harrison, R.F. & Kennedy, R.L. 1997, 'Application of autonomous neural network systems to medical pattern classification tasks', *Artificial Intelligence in Medicine*, vol. 11, no. 3, pp. 215-239.

Lin, Y. & Jeon, Y. 2002, *Random forest and adaptive nearest neighbors*, Technical Report 1055, Department of Statistics, University of Wisconsin, Madison, Wisconsin, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.75.2319&rep=rep1&type=pdf>.

Lindsay, M.D., Jessell, M.W., Ailleres, L., Perrouty, S., de Kemp, E. & Betts, P.G. 2013, 'Geodiversity: exploration of 3D geological model space', *Tectonophysics*, vol. 594, pp. 27-37.

Link, C.A. & Blundell, S. 2003, 'Interpretation of Shallow Stratigraphic facies using a Self-Organizing Neural Network', in W. Sandham & M. Leggett (eds), *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 215-230.

Lloyd, C.D. 2011, *Local Models for Spatial Analysis*, 2nd edn, CRC Press, Boco Raton, Florida, p. 336.

Loosvelt, L., Peters, J., Skriver, H., Lievens, H., Van Coillie, F.M.B., De Baets, B. & Verhoest, N.E.C. 2012, 'Random Forests as a tool for estimating uncertainty at pixel-level in SAR image classification', *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, pp. 173-184.

Lu, D. & Weng, Q. 2007, 'A survey of image classification methods and techniques for improving classification performance', *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823-870.

MacKay, D.J.C. 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, p. 640.

Maiti, S. & Tiwari, R.K. 2010a, 'Automatic discriminations among geophysical signals via the Bayesian neural networks approach', *Geophysics*, vol. 75, no. 1, pp. E67-E78.

Maiti, S. & Tiwari, R.K. 2010b, 'Neural network modeling and an uncertainty analysis in Bayesian framework: A case study from the KTB borehole site', *Journal of Geophysical Research*, vol. 115, no. B10, p. B10208.

Malinverno, A. 2002, 'Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem', *Geophysical Journal International*, vol. 151, no. 3, pp. 675-688.

Marsh, I. & Brown, C. 2009, 'Neural network classification of multibeam backscatter and bathymetry data from Stanton Bank (Area IV)', *Applied Acoustics*, vol. 70, no. 10, pp. 1269-1276.

Marsland, S. 2009, *Machine Learning: An Algorithmic Perspective*, Machine Learning and Pattern Recognition Series, Chapman & Hall/CRC Press, Boca Raton, Florida, p. 406.

Masotti, M., Falsaperla, S., Langer, H., Spampinato, S. & Campanini, R. 2006, 'Application of Support Vector Machine to the classification of volcanic tremor at Etna, Italy', *Geophysical Research Letters*, vol. 33, no. 20, p. L20304.

McCulloch, W.S. & Pitts, W.H. 1943, 'A logical calculus of the ideas immanent in nervous activity', *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115-133.

McNeill, A.W. & Corbett, K.D. 1989, *Geology of the Tullah–Mt Block area*, Tasmanian Department of Mines, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/MRV2/MRV2.pdf>.

McNeill, A.W., de Bomford, R. & Richardson, S.M. 1998, *Relinquishment Report, EL 106/87, Lake Mackintosh*, 98-4116, Mineral Resources Tasmania, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/tasxplor/download/98_4116/98_4116.zip>.

Melgani, F. & Bruzzone, L. 2004, 'Classification of hyperspectral remote sensing images with support vector machines', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778-1790.

Merdith, A.S., Landgrebe, T.C.W., Dutkiewicz, A. & Müller, R.D. 2013, 'Towards a predictive model for opal exploration using a spatio-temporal data mining approach', *Australian Journal of Earth Sciences*, vol. 60, no. 2, pp. 217-229.

Metelka, V., Baratoux, L., Naba, S. & Jessell, M.W. 2011, 'A geophysically constrained litho-structural analysis of the Eburnean greenstone belts and associated granitoid domains, Burkina Faso, West Africa', *Precambrian Research*, vol. 190, no. 1–4, pp. 48-69.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994a, 'Conclusions', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 213-227.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994b, 'Introduction', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 1-5.

Miller, H.J. & Han, J. 2001, 'Geographic Data Mining and Knowledge Discovery: An Overview ', in H.J. Miller & J. Han (eds), *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, London, pp. 3 - 32.

Mitchell, J.M.O. 1994, 'Classical Statistical Methods', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 17-28.

Mineral Resources Tasmanian 2011, *1:25,000 Scale Digital Geology of Tasmania*, Mineral Resources Tasmania, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/>.

Molina, R., P´erez de la Blanca, N. & Taylor, C.C. 1994, 'Modern Statistical Techniques', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 29-49.

Morris, R., Asiedu, J., Haber, W., SaintOurs, F., Stevenson, R. & Tang, H. 2007, 'Database-backed decision trees with application to biological informatics', *Journal of Intelligent Information Systems*, vol. 29, no. 1, pp. 25-38.

Mosegaard, K. & Tarantola, A. 1995, 'Monte Carlo sampling of solutions to inverse problems', *Journal of Geophysical Research*, vol. 100, no. B7, pp. 12431-12431.

Mshiu, E.E. 2011, 'Landsat remote sensing data as an alternative approach for geological mapping in Tanzania: a case study in the Rungwe Volcanic Province, south-western Tanzania', *Tanzania Journal of Science*, vol. 37, pp. 26-36.

Mucke, H.A.M. 2009, 'Making Sense of Data', *Bio - IT World*, vol. 8, no. 5, pp. 10-10.

Mukherjee, M. & Singh, S.B. 2009, 'Artificial Neural Network: some applications in physical metallurgy of steels', *Materials and Manufacturing Processes*, vol. 24, no. 2, pp. 198-208.

Murray, H., Lucieer, A. & Williams, R. 2010, 'Texture-based classification of sub-Antarctic vegetation communities on Heard Island', *International Journal of Applied Earth Observation and Geoinformation*, vol. 12, no. 3, pp. 138-149.

National Aeronautics and Space Administration 2000, *Landsat ETM+ scene ELP096R082_7T20001010, L1G*, Landsat Program, Sioux Falls, South Dakota.

National Aeronautics and Space Administration 2002, *Landsat ETM+ scene ELP090R089_7T20020428, L1G* Landsat Program, Sioux Falls, South Dakota..

National Aeronautics and Space Administration, 2010, *Landsat TM*, Earth Observing System, < http://eospso.gsfc.nasa.gov/>.

NeCTAR 2011, *National eResearch Collaboration Tools and Resources - Home*, Commonwealth of Australia, Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education , Canberra, Australian Capital Territory, <https://www.nectar.org.au>.

Niaf, E., Flamary, R., Lartizien, C. & Canu, S. 2011, 'Handling uncertainties in SVM classification', in *IEEE Statistical Signal Processing Workshop*, June 2011, Nice, France, pp. 757-760.

Noll, C.A. & Hall, M. 2005, 'Structural architecture of the Owen Conglomerate, West Coast Range, western Tasmania: field evidence for Late Cambrian extension', *Australian Journal of Earth Sciences*, vol. 52, no. 3, pp. 411-426.

Oh, H.-J. & Lee, S. 2010, 'Application of Artificial Neural Network for gold–silver deposits potential mapping: a case study of Korea', *Natural Resources Research*, vol. 19, no. 2, pp. 103-124.

Oommen, T., Misra, D., Twarakavi, N.K.C., Prakash, A., Sahoo, B. & Bandopadhyay, S. 2008, 'An objective analysis of support vector machine based classification for remote sensing', *Mathematical Geosciences*, vol. 40, no. 4, pp. 409-424.

Ord, J.K. & Getis, A. 1995, 'Local spatial autocorrelation statistics: distributional issues and an application', *Geographical Analysis*, vol. 27, pp. 286-306.

Paasche, H. & Eberle, D.G. 2009, 'Rapid integration of large airborne geophysical data suites using a fuzzy partitioning cluster algorithm: a tool for geological mapping and mineral exploration targeting', *Exploration Geophysics*, vol. 40, no. 3, pp. 277-287.

Pacifici, F., Chini, M. & Emery, W.J. 2009, 'A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification', *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1276-1292.

Page, R.W., Conor, C.H.H., Stevens, B.P.J., Gibson, G.M., Preiss, W.V. & Southgate, P.N. 2005a, 'Correlation of Olary and Broken Hill Domains, Curnamona Province: Possible relationship to Mt Isa and other North Australian Pb-Zn-Ag-bearing successions', *Economic Geology*, vol. 100, no. 4, pp. 663-676.

Page, R.W., Stevens, B.P.J. & Gibson, G.M. 2005b, 'Geochronology of the sequence hosting the Broken Hill Pb-Zn-Ag orebody, Australia', *Economic Geology*, vol. 100, no. 4, pp. 633-661.

Pal, M. 2005, 'Random forest classifier for remote sensing classification', *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217-222.

Pal, M. & Mather, P.M. 2003, 'An assessment of the effectiveness of decision tree methods for land cover classification', *Remote Sensing of Environment*, vol. 86, no. 4, pp. 554-565.

Pal, M. & Mather, P.M. 2005, 'Support vector machines for classification in remote sensing', *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007-1011.

Park, Y.-S., Céréghino, R., Compin, A. & Lek, S. 2003, 'Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters', *Ecological Modelling*, vol. 160, no. 3, pp. 265-280.

Paruelo, J. & Tomasel, F. 1997, 'Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models', *Ecological Modelling*, vol. 98, no. 2-3, pp. 173-186.

Pearl, J. 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, California, p. 552.

Pebesma, E.J. & Bivand, R.S. 2005, 'Classes and methods for spatial data in R', *R News*, vol. 5, no. 2, pp. 9-13.

Penn, B.S. 2005, 'Using self-organizing maps to visualize high-dimensional data', *Computers & Geosciences*, vol. 31, no. 5, pp. 531-544.

Peters, A. & Hothorn, T. 2011, *ipred: improved predictors, R package version 0.8-11*, <http://CRAN.R-project.org/package=ipred>.

Pichler, M.A. & Perone, S.P. 1974, 'Computerized pattern recognition applications to chemical analysis. Development of interactive feature selection methods for the k-nearest neighbor technique', *Analytical Chemistry*, vol. 46, no. 12, pp. 1790-1798.

Price, D., Knerr, S., Personnaz, L. & Dreyfus, G. 1995, 'Pairwise neural network classifiers with probabilistic outputs', in G. Tesauro, D. Touretzky & T. Leen (eds), *Neural Information Processing Systems*, The MIT Press, Cambridge, Massachusetts, vol. 7, pp. 1109–1116.

Provost, F. & Fawcett, T. 1997, 'Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions', in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, October 1996, Huntington Beach, California, pp. 43-48.

Ramli, M.F., Yusof, N., Yusoff, M.K., Juahir, H. & Shafri, H.Z.M. 2010, 'Lineament mapping and its application in landslide hazard assessment: a review', *Bulletin of Engineering Geology and the Environment*, vol. 69, no. 2, pp. 215-233.

Ramstein, G. & Raffy, M. 1989, 'Analysis of the structure of radiometric remotely-sensed images', *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 1049-1073.

Reid, J.E. 2003, *Western Tasmanian Regional Minerals Program. helicopter electromagnetic data - processing, quality control and interpretation*, UR2003_09 University of Tasmania, Hobart, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/UR2003_09/>.

Reitsma, F. 2010, 'Geoscience explanations: identifying what is needed for generating scientific narratives from data models', *Environmental Modelling & Software*, vol. 25, no. 1, pp. 93-99.

Ricchetti, E. 2000, 'Multispectral satellite image and ancillary data integration for geological classification', *Photogrammetric Engineering and Remote Sensing*, vol. 66, no. 4, pp. 429-435.

Richardson, S.M. 1993, *Exploration Licence, 106/87, Lake Mackintosh, Tasmanian Progress Report for the Period April 1992 to April 1993*, 93-3441, Aberfoyle Resources Limited, Mineral Resources Tasmania, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/tasxplor/download/93_3441/93-3441.pdf>.

Richardson, S.M. 1994, *Exploration Licence 106/87 Lake Mackintosh, Tasmanian Progress Report for the Period April 1993 to February 1994*, 94-3537, Aberfoyle Resources Limited, Mineral Resources Tasmania, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/tasxplor/download/94_3537/94-3537.pdf>.

Ripley, B.D. 1996, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, p. 403.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. 2011, 'pROC: an open-source package for R and S+ to analyze and compare ROC curves', *BMC Bioinformatics*, vol. 12, no. 77, p. 8.

Rogers, S.J., Fang, J.H., Karr, C.L. & Stanley, D.A. 1992, 'Determination of lithology from well logs using a neural network', *AAPG Bulletin*, vol. 76, no. 5, pp. 731-739.

Rohwer, R., Wynne-Jones, M. & Wysotzki, F. 1994, 'Neural Networks', in D. Michie, D.J. Spiegelhalter & C.C. Taylor (eds), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, pp. 84-106.

Rojas, R. 1996, *Neural Networks: A Systematic Introduction*, Springer-Verlag, Berlin, p. 502.

Rosenblatt, F. 1962, *Principles of Neurodynamics*, Spartan Books, Washington, DC, p. 616.

Rubenach, M.J. 1974, 'The origin and emplacement of the Serpentine Hill Complex, Western Tasmania', *Journal of the Geological Society of Australia*, vol. 21, no. 1, pp. 91-106.

Sabatakakis, N., Koukis, G., Vassiliades, E. & Lainas, S. 2012, 'Landslide susceptibility zonation in Greece', *Natural Hazards*, vol. 65, no. 1, pp. 523-543.

Sambridge, M. 1999a, 'Geophysical inversion with a neighbourhood algorithm-I. Searching a parameter space', *Geophysical Journal International*, vol. 138, no. 2, pp. 479-494.

Sambridge, M. 1999b, 'Geophysical Inversion with a neighbourhood algorithm - II. Appraising the ensemble', *Geophysical Journal International*, vol. 138, pp. 727-749.

Sambridge, M., Beghein, C., Simons, F. & Snieder, R. 2006, 'How do we understand and visualize uncertainty?', *The Leading Edge*, vol. 25, no. 5, pp. 542 - 546.

Savu-Krohn, C., Rantitsch, G., Auer, P., Melcher, F. & Graupner, T. 2011, 'Geochemical fingerprinting of coltan ores by machine learning on uneven datasets', *Natural Resources Research*, vol. 20, no. 3, pp. 177-191.

Scales, A.J. & Snieder, R. 1998, 'What is noise?', *Geophysics*, vol. 63, no. 4, pp. 1122-1124.

Schölkopf, B. 2003, 'Statistical learning theory, capacity, and complexity', *Complexity*, vol. 8, no. 4, pp. 87-94.

Schumann, A. 1997, 'Neural networks versus statistics; a comparing study of their classification performance on well log data', in *Proceedings of the 3rd Annual Conference of the International Association for Mathematical Geology*, September 1997, Barcelona, Spain, vol. 3, pp. 237-241.

Scott, K.M. & Pain C.F. 2008, '*Regolith Science*', CSIRO Publishing, Collingwood, Australia, p. 461.

Segata, N. & Blanzieri, E. 2009, 'Fast Local Support Vector Machines for Large Datasets', in P. Perner (ed.), *Machine Learning and Data Mining in Pattern Recognition*, Springer-Verlag Berlin, Germany, vol. 5632, pp. 295-310.

Segata, N. & Blanzieri, E. 2010, 'Fast and scalable local kernel machines', *The Journal of Machine Learning Research*, vol. 11, pp. 1883-1926.

Segata, N., Pasolli, E., Melgani, F. & Blanzieri, E. 2012, 'Local SVM approaches for fast and accurate classification of remote-sensing images', *International Journal of Remote Sensing*, vol. 33, no. 19, pp. 6186-6201.

Sellars, S., Nguyen, P., Chu, W., Gao, X., Hsu, K.-l. & Sorooshian, S. 2013, 'Computational earth science: Big Data transformed into insight', *Eos, Transactions American Geophysical Union*, vol. 94, no. 32, pp. 277-278.

Sen, M.K. & Stoffa, P.L. 1992, 'Rapid sampling of model space using genetic algorithms - examples from seismic waveform inversion', *Geophysical Journal International*, vol. 108, no. 1, pp. 281-292.

Seymour, D.B. & Calver, C.R. 1995, *Explanatory notes for the Time–Space Diagram and Stratotectonic Elements Map of Tasmania*, Tasmanian Geological Survey Record 1995/01, Tasmanian Geological Survey, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/UR1995_01/UR1995_01.pdf>.

Seymour, D.B., Green, G.R. & Calver, C.R. 2013, *The Geology and Mineral Deposits of Tasmania: a summary*, Mineral Resources Tasmania, Department of Infrastructure, Energy and Resources, Rosny Park, Tasmania, <http://www.mrt.tas.gov.au/mrtdoc/dominfo/download/GSB72_3/bulletin_72_3.pdf>.

Shaheen, M., Shahbaz, M., Rehman, Z. & Guergachi, A. 2011, 'Data mining applications in hydrocarbon exploration', *Artificial Intelligence Review*, vol. 35, no. 1, pp. 1-18.

Shankar, V. 2009, *Texture-Based Automated Lithological Classification using Aeromagnetic Anomaly Images*, Master of Science, Department of Electrical and

Computer Engineering University of Arizona, Tucson, Arizona, p. 136, <http://pubs.usgs.gov/of/2009/1206/of2009-1206.pdf>.

Shi, H. & Liu, Y. 2011, 'Naïve Bayes vs. Support Vector Machine: Resilience to Missing Data', in H. Deng, D. Miao, J. Lei & F.-L. Wang (eds), *Artificial Intelligence and Computational Intelligence*, Springer, Berlin, Germany, vol. 7003, pp. 680-687.

Shin, K.-S., Lee, T.S. & Kim, H.-j. 2005, 'An application of support vector machines in bankruptcy prediction model', *Expert Systems with Applications*, vol. 28, no. 1, pp. 127-135.

Shmueli, G. 2010, 'To explain or to predict?', *Statistical Science*, vol. 25, no. 3, pp. 289-310.

Sinclair , B.J. 1994, 'Geology and Geochemistry of the Que River Shale, Western Tasmania', B.Sc. (Hons.), School of Earth Sciences, University of Tasmania, Hobart, Tasmania, p. 114.

Singer, D. & Kouda, R. 1996, 'Application of a feedforward neural network in the search for Kuroko deposits in the Hokuroku district, Japan', *Mathematical Geology*, vol. 28, no. 8, pp. 1017-1023.

Sivia, D.S. 1996, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, New York, p. 189.

Skabar, A. 2007, 'Modeling the spatial distribution of mineral deposits using neural networks', *Natural Resource Modeling*, vol. 20, no. 3, pp. 435-450.

Sklavounos, P. & Sakellariou, M. 1995, 'Intelligent classification of rock masses', *Transactions on Information and Communications Technologies*, vol. 8, pp. 387-393.

Slavinski, H., Morris, B., Ugalde, H., Spicer, B., Skulski, T. & Rogers, N. 2010, 'Integration of lithological, geophysical, and remote sensing information: a basis for remote predictive geological mapping of the Baie Verte Peninsula, Newfoundland', *Canadian Journal of Remote Sensing*, vol. 36, no. 2, pp. 99-118.

Smirnoff, A., Boisvert, E. & Paradis, S.J. 2008, 'Support vector machine for 3D modelling from sparse geological information of various origins', *Computers & Geosciences*, vol. 34, no. 2, pp. 127-143.

Song, C., Woodcock, C.E., Seto, K.C., Lenney, M.P. & Macomber, S.A. 2001, 'Classification and change detection using Landsat TM Data: when and how to correct atmospheric effects?', *Remote Sensing of Environment*, vol. 75, no. 2, pp. 230-244.

Song, X., Duan, Z. & Jiang, X. 2012, 'Comparison of artificial neural networks and support vector machine classifiers for land cover classification in Northern China using a SPOT-5 HRG image', *International Journal of Remote Sensing*, vol. 33, no. 10, pp. 3301-3320.

Spencer, C. & Gubbins, D. 1980, 'Travel-time inversion for simultaneous earthquake location and velocity structure determination in laterally varying media', *Geophysical Journal of the Royal Astronomical Society*, vol. 63, no. 1, pp. 95-116.

Statnikov, A., Wang, L. & Aliferis, C.F. 2008, 'A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification', *BMC Bioinformatics*, vol. 9, no. 1, p. 319.

Stepinski, T.F. & Bue, B.D. 2006, 'Automated classification of landforms on Mars', *Computers & Geosciences*, vol. 32, no. 5, pp. 604-614.

Stepinski, T.F., Ghosh, S. & Vilalta, R. 2007, 'Machine learning for automatic mapping of planetary surfaces', in *Proceedings of the 21st National Conference on Artificial Intelligence*, July 2007, Vancouver, British Columbia, vol. 2, pp. 1807-1812.

Stevens, B.P.J. 1986, 'Post-depositional history of the Willyama Supergroup in the Broken Hill Block, NSW', *Australian Journal of Earth Sciences*, vol. 33, no. 1, pp. 73-98.

Stevens, B.P.J., Willis, I.L., Brown, R.E. & Stroud, W.J. 1983, 'The Early Proterozoic Willyama Supergroup: definitions of stratigraphic units from the Broken Hill Block, New South Wales', *Geological Survey of New South Wales, Records*, vol. 21, no. 2, pp. 407-442.

Stow, D. 2010, 'Geographic Object-Based Image Change Analysis', in M.M. Fisher & A. Getis (eds), *Handbook of Applied Spatial Analysis: Software, Tools, Methods and Applications*, Springer-Verlag, Berlin, Germany, pp. 565-582.

Stumpf, A. & Kerle, N. 2011, 'Object-oriented mapping of landslides using Random Forests', *Remote Sensing of Environment*, vol. 115, no. 10, pp. 2564-2577.

Tan, P.-N., Steinbach, M. & Kumar, V. 2006, *Introduction to Data Mining*, Addison-Wesley, Boston, Massachusetts, p. 769.

Tangestani, M.H. 2004, 'Landslide susceptibility mapping using the fuzzy gamma approach in a GIS, Kakan catchment area, southwest Iran', *Australian Journal of Earth Sciences*, vol. 51, no. 3, pp. 439-450.

Tarabalka, Y., Benediktsson, J.A. & Chanussot, J. 2009, 'Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 8, pp. 2973-2987.

Taye, W. 2011, 'Lithology boundary detection using multi-sensor remote sensing imagery for geological interpretation', Master of Science, Faculty or Geo-Information Science and Earth Observation, University of Twente, Enscheded, Netherlands, <http://www.itc.nl/library/papers_2011/msc/aes/taye.pdf>.

Telford, W.M., Geldart, L.P. & Sheriff, R.E. 1990, *Applied Geophysics*, 2nd edn, Cambridge University Press, Cambridge, p. 770.

Thomas, W.A. 2004, *Meeting Challenges with Geologic Maps*, American Geological Institute, Alexandria, Virginia, < http://www.agiweb.org/environment/publications/mapping/mappingbook.pdf>.

Tobler, W.R. 1970, 'A computer movie simulating urban growth in the Detroit region', *Economic Geography*, vol. 46, no. 2, pp. 234-240.

Tomes, K.L. 2011, 'The Textures and Geochemistry of the Hellyer Basalt, western Tasmania', B.Sc. (Hons.), School of Earth Sciences, University of Tasmania, Hobart, Tasmanian, p. 91.

Toumani, A. 2003, 'Fuzzy Classification for Lithology Determination from Well Logs ', in W. Sandham & M. Leggett (eds), *Geophysical Applications of Artificial Neural Networks and Fuzzy Logic*, Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 125-142.

Trustorff, J.-H., Konrad, P. & Leker, J. 2011, 'Credit risk prediction using support vector machines', *Review of Quantitative Finance and Accounting*, vol. 36, no. 4, pp. 565-581.

Turner, N.J. 1989, 'Precambrian', in C.F. Burrett & E.L. Martin (eds), *Geology and Mineral Resources of Tasmania*, Special Publication, Geological Society of Australia, Brisbane, Queensland, vol. 15, pp. 5-46.

Turner, N.J., Black, L.P. & Kamperman M., 1998, 'Dating of Neoproterozoic and Cambrian orogenies in Tasmania', *Australian Journal of Earth Sciences*, vol. 45, no. 5, p. 789.

Ultsch, A. & Vetter, C. 1994, *Self-organising Feature Maps versus Statistical Clustering A Benchmark*, Department of Mathematics and Computer Science, University of Marburg, Germany, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.196.7322&rep=rep1&type=pdf>.

UNCOVER 2012, *Searching the Deep Earth: A Vision for Exploration Geoscience in Australia*, Australian Academy of Science, Canberra, Australian Capital Territory, <http://www.science.org.au/policy/uncover.html/>.

Uriarte, E.A. & Martín, F.D. 2005, 'Topology preservation in SOM', *International Journal of Applied Mathematics and Computer Sciences*, vol. 1, no. 1, pp. 19-22.

van der Baan, M. & Jutten, C. 2000, 'Neural networks in geophysical applications', *Geophysics*, vol. 65, no. 4, pp. 1032-1047.

van der Wel, F.J.M., van der Gaag, L.C. & Gorte, B.G.H. 1998, 'Visual exploration of uncertainty in remote-sensing classification', *Computers & Geosciences*, vol. 24, no. 4, pp. 335-343.

Vapnik, V.N. 1995, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, p. 188.

Vapnik, V.N. 1998, *Statistical Learning Theory*, Wiley, New York, p. 736.

Venables, W.N. & Ripley, B.D. 2002, *Modern Applied Statistics with S*, 4th edn, Statistics and Computing, Springer, New York, p. 495.

Vesanto, J. 1999, 'SOM-based data visualization methods', *Intelligent Data Analysis*, vol. 3, no. 2, pp. 111-126.

Vesanto, J. & Alhoniemi, E. 2000, 'Clustering of the self-organizing map', *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586-600.

Wagstaff, K. & Bell, J.F. 2003, 'Automated analysis of mars multispectral observations', in *The 6th International Conference on Mars*, July 2003 Pasadena, California, abstract no.3120.

Wang, S. & Yao, X. 2012, 'Multiclass imbalance problems: analysis and potential solutions', *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 4, pp. 1119-1130.

Wang, X., Niu, R. & Wu, K. 2011, 'Lithology intelligent identification using support vector machine and adaptive cellular automata in multispectral remote sensing image', *Optical Engineering*, vol. 50, no. 7.

Waske, B., Benediktsson, J.A., Árnason, K. & Sveinsson, J.R. 2009, 'Mapping of hyperspectral AVIRIS data using machine-learning algorithms', *Canadian Journal of Remote Sensing*, vol. 35, no. 1, pp. 106-116.

Waske, B., Benediktsson, J.A. & Sveinsson, J.R. 2012, 'Random Forests Classification of Remote Sensing Data', in C.H. Chen (ed.), *Signal and Image Processing for Remote Sensing*, 2nd edn, CRC Press, Hoboken, New Jersey, pp. 365-374.

Waske, B. & Braun, M. 2009, 'Classifier ensembles for land cover mapping using multitemporal SAR imagery', *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 5, pp. 450-457.

Waske, B., Van Der Linden, S., Benediktsson, J.A., Rabe, A. & Hostert, P. 2010, 'Sensitivity of support vector machines to random feature selection in classification of hyperspectral data', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 7, pp. 2880-2889.

Waters, J.C. & Wallace, D.B. 1992, 'Volcanolgy and sedimentology of the host sequence to the Hellyer and Que River volcanic-hosted massive sulfide deposits, northwestern Tasmania', *Economic Geology*, vol. 87, no. 3, pp. 650-666.

Webb, G.I. 2000, 'MultiBoosting: A technique for combining boosting and wagging', *Machine Learning*, vol. 40, no. 2, pp. 159-196.

Webster, A.E. 2004, 'The structural evolution of the Broken Hill Pb-Zn-Ag deposit, New South Wales, Australia', Ph.D., School of Earth Sciences, University of Tasmania, Hobart, Tasmania, p. 430.

Wehrens, R. & Buydens, L.M.C. 2007, 'Self- and super-organising maps in R: the kohonen package', *Journal of Statistical Software*, vol. 21, no. 5, pp. 1-19.

Weihs, C., Ligges, U., Luebke, K. & Raabe, N. 2005, 'klaR Analyzing German Business Cycles', in D. Baier, R. Decker & L. Schmidt-Thieme (eds), *Data Analysis and Decision Support*, Springer-Verlag, Berlin, Germany, pp. 335-343.

Wellmann, J.F. 2011, 'Uncertainties have a Meaning: Quantitative Interpretation of the Relationship between Subsurface Flow and Geological Data Quality', Ph.D., School of Earth and Environment, University of Western Australia Perth, Western Australia, p. 195.

Wellmann, J.F. & Regenauer-Lieb, K. 2012, 'Uncertainties have a meaning: Information entropy as a quality measure for 3-D geological models', *Tectonophysics*, vol. 526–529, pp. 207-216.

Wettschereck, D., Aha, D. & Mohri, T. 1997, 'A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms', *Artificial Intelligence Review*, vol. 11, no. 1-5, pp. 273-314.

Williams, D. 2009, *Landsat 7 Science Data user's Handbook*, National Aeronautics and Space Administration, Greenbelt, Maryland, p. 186.

Willis, I.L., Brown, R.E., Stroud, W.J. & Stevens, B.P.J. 1983, 'The early Proterozoic Willyama supergroup: stratigraphic subdivision and interpretation of high to low grade metamorphic rocks in the Broken Hill Block, New South Wales', *Journal of the Geological Society of Australia*, vol. 30, no. 1-2, pp. 195-224.

Wilson, M.D., Ustin, L. & Rocke, D.M. 2004, 'Classification of contamination in salt marsh plants using hyperspectral reflectance', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, pp. 1088-1095.

Witten, I.H. & Frank, E. 2005, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn, Morgan Kaufmann Series in Data Management Systems, Elsevier/Morgan Kaufman, San Francisco, California, p. 525.

Wong, P.M., Jian, F.X. & Taggart, I.J. 1995, 'A critical comparison of neural networks and discriminant analysis in lithofacies, porosity and permeability predictions', *Journal of Petroleum Geology*, vol. 18, no. 2, pp. 191-206.

Wu, C.-F., Lin, C.-J. & Lee, C.-Y. 2011, 'A functional neural fuzzy network for classification applications', *Expert Systems with Applications*, vol. 38, no. 5, pp. 6202-6208.

Wu, T.-F., Lin, C.-J. & Weng, R.C. 2004, 'Probability estimates for multi-class classification by pairwise coupling', *Journal of Machine Learning Research*, vol. 5, pp. 975-1005.

Xie, X.L. & Beni, G. 1991, 'A validity measure for fuzzy clustering', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 841–847.

Xu, R. & Wunsch, D. 2005, 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678.

Ya, X., Xuejun, L., Carin, L. & Krishnapuram, B. 2007, 'Multi-Task learning for classification with Dirichlet process priors', *Journal of Machine Learning Research*, vol. 8, no. 1, pp. 35-63.

Yan, Z., Li, J.G., Xiong, Y.M., Xu, W.T. & Zheng, G.R. 2012, 'Identification of candidate colon cancer biomarkers by applying a random forest approach on microarray data', *Oncology Reports*, vol. 28, no. 3, pp. 1036-1042.

Yang, F., Wang, H.-z., Mi, H., Lin, C.-d. & Cai, W.-w. 2009, 'Using random forest for reliable classification and cost-sensitive learning for medical diagnosis', *BMC Bioinformatics*, vol. 10, no. 1, p. S22.

Yang, G., Collins, M.J. & Gong, P. 1998, 'Multisource data selection for lithologic classification with artificial neural networks', *International Journal of Remote Sensing*, vol. 19, no. 18, pp. 3675-3680.

Yu, L. & Liu, H. 2004, 'Efficient feature selection via analysis of relevance and redundancy', *Journal of Machine Learning Research*, vol. 5, pp. 1205-1224.

Yu, L., Porwal, A., Holden, E.J. & Dentith, M.C. 2012, 'Towards automatic lithological classification from remote sensing data using support vector machines', *Computers & Geosciences*, vol. 45, pp. 229-239.

Zammit, O., Descombes, X. & Zerubia, J. 2007, 'Assessment of different classification algorithms for burnt land discrimination', in *IEEE International Geoscience and Remote Sensing Symposium*, July 2007, Barcelona, Spain, pp. 3000-3003.

Zhang, H., Berg, A.C., Maire, M. & Malik, J. 2006, 'SVM-KNN: discriminative nearest neighbor classification for visual category recognition', in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2006, New York, vol. 2, pp. 2126-2136.

Zhou, X. & Chen, Y. 2005, 'Seafloor classification of multibeam sonar data using neural network approach', *Marine Geodesy*, vol. 28, no. 2, pp. 201-206.

Zhu, A.X. 1997, 'Measuring uncertainty in class assignment for natural resource maps under fuzzy logic', *Photogrammetric Engineering and Remote Sensing*, vol. 63, no. 10, pp. 1195-1202.

Zortea, M., De Martino, M. & Serpico, S. 2007, 'A SVM ensemble approach for spectral-contextual classification of optical high spatial resolution imagery', in *IEEE*

International Geoscience and Remote Sensing Symposium, July 2007, Barcelona, Spain, pp. 1489-1492.

APPENDIX A – MACHINE LEARNING ALGORITHM SENSITIVITY TO IMBALANCED CLASS DISTRIBUTIONS

Imbalanced distributions of classes in labelled samples are often encountered real-world supervised classification applications. Imbalanced (skewed) class distributions are reported to have a varying degree of impact on the predictions generated by machine learning algorithms. This document details a series of experiments designed to assess the sensitivity of five different machine learning algorithms to imbalanced training data class distributions when faced with random input variables. The results of this study are presented in terms of the *p*-value. The *p*-value provides an indication of the probability that predictions are significantly better than a random guess. The *k*-Nearest Neighbours and Random Forests machine learning algorithms, unlike Naïve Bayes, Support Vector Machines or Artificial Neural Networks, are observed to be relatively insensitive to imbalanced training data. These findings suggest that inducing multiple, locally (in variable space) constrained classifiers reduces the impact of class imbalance. In contrast, assessing all training samples together using a single global classifier will generate skewed results when trained on imbalance class distributions and noisy input data.

A.1. Introduction

The performances of five machine learning algorithms (MLAs) are evaluated in the context of a supervised image classification task utilising input data containing random values. Evaluations are assessed with respect to the number of classes and imbalanced class distributions of both training (T_a) and test data (T_b). Imbalanced class distributions, where one class is represented by a comparatively large number of samples compared to the other classes, are often encountered in real-world applications (Henery 1994b; Wang & Yao 2012). Class imbalance, or skewed class distributions within labelled instances T, used to sample T_a and T_b, is reported to have a significant effect on the results of MLAs such as Decision Trees and to a lesser extent, Artificial Neural Networks (Japkowicz & Stephen 2002).

In this study, the effect of imbalanced class distributions in both T_a and T_b on MLA prediction accuracy is evaluated given random input data. Random input data is used as a means of assessing how MLAs respond to noisy input variables, a situation commonly encountered in geophysical applications (Scales & Snieder 1998). The hypothesis for this experiment is that all MLAs will respond in a similar fashion to imbalanced T_a using random inputs by defaulting to the classification of the most common class within these data. Therefore, MLA prediction accuracy will correspond to the proportion of the most common class in the T_a, i.e. the "no-information" rate (Kuhn *et al.* 2012). The *p*-value is used here to quantify the probability that MLA predictions formulated using real-world data from a supervised image classification task will be significantly more accurate than a random guess. The *p*-value represents $p(\mathrm{X} > x)$, where X is the expected accuracy given "no-information" and x is the observed accuracy.

A.2. Methods

Five MLAs are compared in this experiment: Naïve Bayes (NB; John & Langley 1995); *k*-Nearest Neighbours (kNN, Fix & Hodges 1951); Random Forests (RF, Breiman 2001); Support Vector Machines (SVM, Vapnik 1995; Vapnik 1998); and Artificial Neural Networks (ANN, Ripley 1996). All processing was conducted in the statistical programming language R using the *raster* package (Hijmans & van Etten 2012) for image processing and manipulation and the *caret* package (Kuhn *et al.* 2012) for MLA training and evaluation.

Random predictions were generated using a single predictive variable, in this case an image (128 × 128 pixels) where each pixel/sample is given a uniformly distributed random value between 0 and 1 (Figure A.1). MLA models were trained using a randomly selected subset of ~ 1000 T_a samples and class predictions were made for the entire image space (T_b). MLA prediction performance is defined using the following measure of overall accuracy,

$$accuracy = \frac{number\ of\ correct\ Tb\ predictions}{total\ number\ of\ Tb\ samples}. \qquad [A.E.1]$$

The *p*-values for observed accuracies between 0 and 1 (at 0.01 increments) were generated by subtracting the cumulative sum of normalised probability densities from 1.

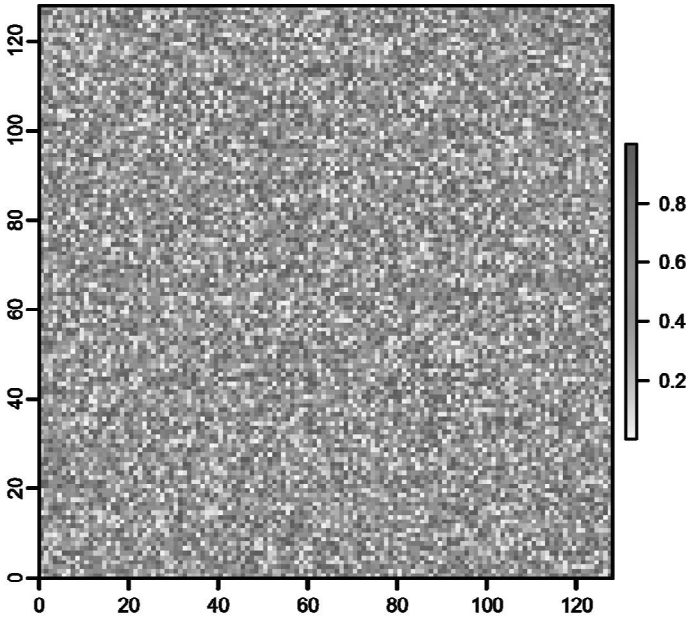

Figure A.1 Image (input variable) used for MLA supervised classification training and testing. Random values between 0 and 1 are assigned using a uniform distribution.

Table A.1 MLA model parameter values used for "no-information" accuracy prediction (default if possible).

MLA	Model parameters	Model parameter values
NB	*usekernel* (parametric/non-parametric density est.)	FALSE
kNN	*k* (number of neighbours)	1
RF	*mtry* (number of variables to sample at each split)	1
	trees (number of trees)	500
SVM	*kernel* (kernel type for dimensionality reduction)	Radial Basis Function
	ó (kernel width)	0.01
	C (support vector misclassification cost)	1
ANN	*decay* (weight decay)	0
	size (units in hidden layer)	1

All MLA models were trained using default model parameters (Table A.1). MLA model training and testing was repeated 1000 times and the resulting density distribution of prediction accuracy normalised. The number of classes, *c*, for a series of tests was set to two, three and six. Equal and unequal class proportions for both T_a and T_b were varied in four ways for the different *c* (Table A.2 and Figure A.2). For all *c* class proportions the equal class distribution was calculated by dividing 1 by *c*. Unequal class proportions for *c* = 2 and *c* = 3, were set such that they contained one dominant class with a proportion 50 % of the total number of samples. For *c* = 6, unequal class proportions were set using class distributions occurring in a real-world example. Table A.3 shows the equal and unequal class distributions used for all experiments with different *c* for T_a and T_b proportions.

Table A.2 Training and validation class distribution combinations for trials 1–4.

Trials	Class distribution combinations (T_a : T_b)
1	equal : equal
2	equal : unequal
3	unequal : equal
4	unequal : unequal

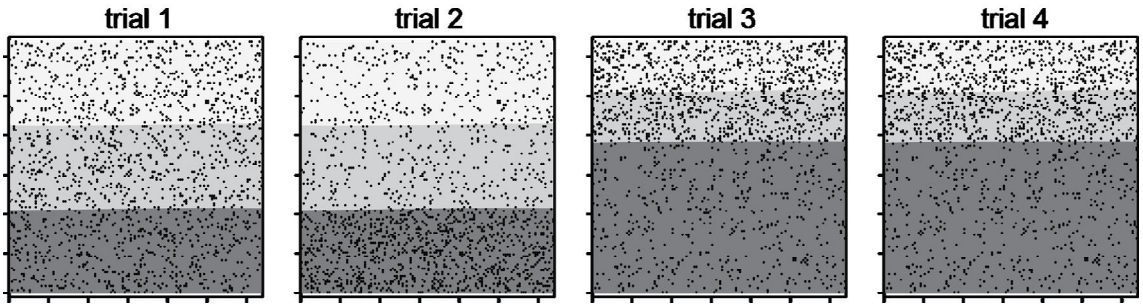

Figure A.2 Three class example of T_a random sampling (points) and T_b (background) sample structure for trials 1–4 (see Tables A.2 and A.3 for details on the distribution of samples). Points represent T_a sample locations and the background colours indicate the distribution of the three classes in T_b.

Table A.3 Equal and unequal class distributions for c = 2, 3 and 6 classes, the unequal distributions for the 6 class prediction task is taken from the distributions found in real-world data.

Classes	Equal class proportions	Unequal class proportions
2	(0.5, 0.5)	(0.4, 0.6)
3	(0.33, 0.33, 0.34)	(0.2, 0.2, 0.6)
6	(0.167, 0.167, 0.167, 0.167, 0.167, 0.167)	(0.49, 0.044, 0.133, 0.253, 0.0404, 0.0394)

A.3. Results

Table A.4 summarises the T_b accuracy results obtained by the five MLAs trialled across all the experiments conducted. For predictions generated using equal class distributions (trial 1) there is little difference between the five MLAs regardless of the number of classes. In all cases, the MLAs have generated a mean prediction accuracy approximating the equal class proportion with standard deviations < 0.005. Trial 2 results, obtained using equal T_a and unequal T_b class distributions, show an increase in accuracy standard deviations for NB, SVM and ANN. This trend is amplified for larger number of classes. This observed trend is due to the increased difference between the equal class proportion and the dominant class unequal class proportion. The mean accuracy given "no-information" for the trial 3 results across all MLAs is equivalent to the equal class proportion, i.e. represented by samples in T_b. Trial 3 results indicate that NB, SVM and ANN generate

Table A.4 MLA T_b accuracy mean ± one standard deviation (n = 1000) across four trials (Table A.2) of equal and unequal class distributions (Table A.3) for two, three and six classes.

	Trial	NB	kNN	RF	SVM	ANN
2 classes	1	0.500 ± 0.002	0.500 ± 0.004	0.500 ± 0.004	0.500 ± 0.002	0.500 ± 0.002
	2	0.499 ± 0.020	0.500 ± 0.005	0.500 ± 0.004	0.500 ± 0.012	0.499 ± 0.040
	3	0.499 ± 0.000	0.500 ± 0.005	0.500 ± 0.004	0.500 ± 0.000	0.499 ± 0.000
	4	0.600 ± 0.000	0.537 ± 0.004	0.520 ± 0.004	0.600 ± 0.000	0.600 ± 0.002
3 classes	1	0.335 ± 0.004	0.333 ± 0.004	0.333 ± 0.004	0.339 ± 0.003	0.335 ± 0.004
	2	0.414 ± 0.064	0.338 ± 0.006	0.336 ± 0.005	0.554 ± 0.074	0.427 ± 0.115
	3	0.340 ± 0.000	0.338 ± 0.003	0.336 ± 0.003	0.340 ± 0.000	0.340 ± 0.000
	4	0.600 ± 0.000	0.508 ± 0.006	0.440 ± 0.006	0.600 ± 0.000	0.600 ± 0.001
6 classes	1	0.166 ± 0.003	0.167 ± 0.003	0.167 ± 0.003	0.167 ± 0.002	0.167 ± 0.002
	2	0.169 ± 0.074	0.167 ± 0.006	0.166 ± 0.005	0.169 ± 0.095	0.169 ± 0.111
	3	0.167 ± 0.000	0.167 ± 0.002	0.167 ± 0.003	0.167 ± 0.000	0.167 ± 0.000
	4	0.490 ± 0.001	0.365 ± 0.006	0.306 ± 0.005	0.490 ± 0.000	0.490 ± 0.002

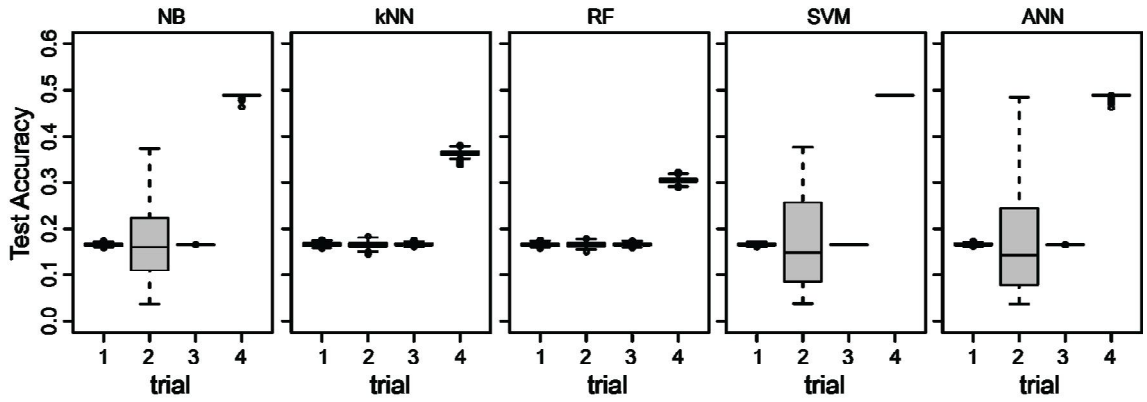

Figure A.3 MLA prediction accuracy distribution boxplots for six classes given "no-information" with four combinations of T_a and T_b class distributions (Table A.2).

negligible T_b accuracy standard deviations, indicating that the only the dominant class has been predicted by these MLAs. The results for trial 4 indicate that the dominant class was predicted for all models generated by NB, SVM and ANN. In contrast, RF and kNN have mean prediction accuracies between the equal class proportion and the dominant class proportion. This suggests that NB, SVM and ANN, when faced with information that does not discriminate between classes, will generate results with distributions mimicking T_a class distributions. In contrast, RF and kNN are more conservative and do not default to predicting only the dominant T_a class for all T_b.

The boxplots in Figure A.3 show the distribution of MLA prediction accuracy for c = 6. The gross trends observed between MLAs for a given trial is identical for the experiments conducted using different numbers of classes. These plots indicate that kNN and RF are

more consistent, with respect to predicted class distributions, when provided with unequal T_a class distributions. This is most evident in trial 2, where kNN and RF present very narrow T_b accuracy distributions. In contrast, NB, SVM and ANN all result in broad trial 2 T_b accuracy distributions. These results imply that NB, SVM and ANN predictions are sensitive to imbalanced T_a class distributions and that T_b accuracy is a function of the random sampling of T_b data. The only other major difference observed between MLAs is in the results of trial 4. In this case, kNN and RF display marginally broader distributions of T_b accuracy with means significantly lower than those obtained from NB, SVM and ANN. This observation confirms that NB, SVM and ANN are all defaulting to the prediction of the dominant T_a class. Conversely, RF and to a lesser extent kNN are consistently generating predictions with balanced class distributions.

Figure A.4 presents normalised probability density distributions generated by the five MLAs evaluated for trial 2 class distributions and c = 6. Superimposed on the density distributions are p-values, represented by the thick black line. It is evident from these curves (and the boxplots in Figure A.3) that the probability that NB, SVM and ANN predictions are random is approaching 0 for T_b accuracies is greater than the dominant class proportion. Hence, the "no-information" rate for $p < 0.05$ (dashed grey line), i.e. probability of 0.95 that predictions are significantly more accurate than a random guess, is marginally less than the proportion of the dominant class in T_b. In contrast, kNN and RF both obtain $p < 0.05$ for T_b accuracies slightly more than the evenly distributed class proportions.

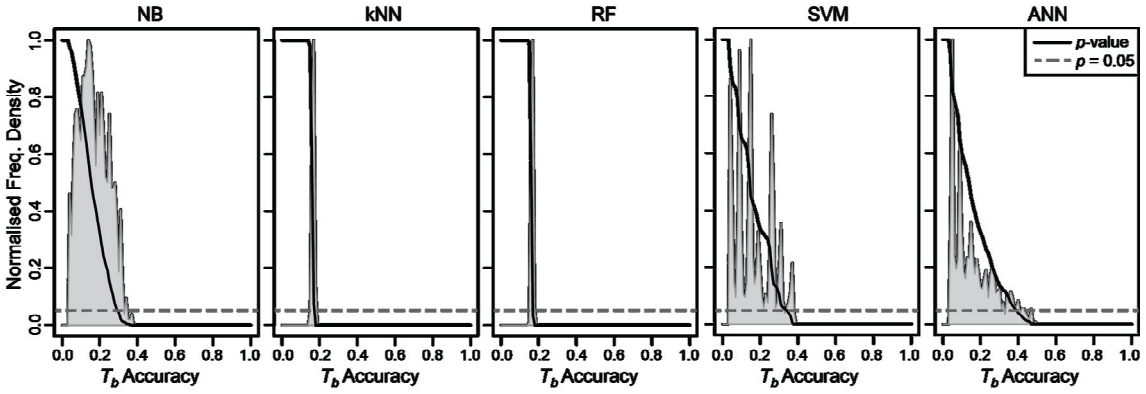

Figure A.4 Trial 2 MLA normalised T_b accuracy density distributions and associated p-values (black lines) for c = 6. T_a class proportions are equal, i.e. 0.1667, T_b class proportions are equivalent to those presented in Table A.3.

A.4. Discussion and Conclusions

When faced with unequal T_a class proportions and random input data, NB, SVM and ANN will default to generating predictions with equivalent class proportions to those represented in T_a. This indicates that NB, SVM and ANN assume that T_a class distributions represent *a priori* information on the expected distributions of predicted class distributions when there is no other information is available to constrain classifier training. This is true for NB in all situations regardless of the information within the input variables (Hastie *et al.* 2009). In contrast, when faced with equal T_a class proportions and random input data, NB, SVM and ANN will generate predictions with highly variable T_a class proportions. In comparison, kNN and RF generate predicted class distributions that are relatively equal for all classes. This observation suggests that kNN and RF generate relatively more conservative predictions than the other MLAs trialled when faced with imbalanced class distributions and "no-information".

The *p*-value represents probability that classifier accuracy is equivalent to that of the "no-information rate", i.e. the expected accuracy of a classifier when randomly generating predictions. Therefore, as the *p*-value approaches 0 the probability that a classifier is more accurate than a random guess is approaching 1. Figure A.4 shows that both kNN and RF generate very narrow T_b accuracy distributions centred on the equal class proportion despite being supplied with imbalanced T_a class distributions. This observation implies that kNN and RF classifiers, induced from imbalanced T_a class distributions, generate predictions that are significantly more accurate than a random guess for comparatively lower accuracies than classifiers induced using NB, SVM or ANN.

The observed differences between kNN and RF, and NB, SVM and ANN may be attributed to the fundamentally different learning strategies employed by these MLAs. The predictions generated by kNN and RF classifiers are based on the combination of multiple classifiers. For example, kNN is an instance-based learner (Kotsiantis 2007), it generates a prediction for a single sample using a majority vote based on the evaluation of neighbouring T_a samples in variable space (Hastie *et al.* 2009). Similarly, RF employs an ensemble strategy that constructs multiple classifiers via bagging. RF class predictions are generated by combining these classifiers using a majority vote (Breiman 2001). The learning strategy employed by RF is similar to that of an adaptive-nearest-Neighbours classifier (Lin & Jeon 2002; Hastie *et al.* 2009). This indicates that these algorithms

generate predictions based on multiple classification models that utilise a subset of locally (in variable space) relevant T_a samples, thus, limiting the bias associated with unequal class distributions. In contrast, NB, SVM and ANN discriminate between classes by constructing a single global classifier (Bernatzki *et al.* 1996; Karatzoglou *et al.* 2006; Hastie *et al.* 2009). Global classifiers, when faced with information that does not discriminate between classes, coupled with unequal T_a class distributions appear to construct decision boundaries that predict only the most abundant class.

The experimental findings presented above indicate that, unlike kNN or RF, MLAs such as NB, SVM and ANN are sensitive to imbalanced T_a class distributions when supplied with random input data for training. This implies that class weighted cost functions need to be employed when training NB, SVM and ANN classifiers on T_a with imbalanced class distributions, or equivalent class distributions are present within T_b to provide an accurate indication of the success of these MLAs.

APPENDIX B – VARIANCE AND ENTROPY FOR MULTICLASS CLASSIFICATION UNCERTAINTY

The outputs of machine learning algorithm supervised classifiers contain c classes, the number of which is determined by the number of classes within labelled data T, used for training and testing. Most machine learning algorithms have the option to output, in conjunction with a class label, a vector of class membership probabilities, p_c, of length equal to the number of classes in T. By default, p_c are normalised such that the sum of its components is equal to 1.

Two metrics, *Variance* and *Entropy*, are commonly used to quantify the degree to which the distribution of p_c is concentrated within a particular class (Goodchild *et al.* 1994; Zhu 1997; Brown 1998; van der Wel *et al.* 1998). *Variance* (Kohavi & Wolpert 1996), a variant of which is called the quadratic score (Glasziou & Hilden 1989), in its normalised form is given by:

$$Variance = \frac{1 - \sum {p_c}^2}{1 - \sum \left(\frac{1}{c}\right)^2}. \qquad [B.1]$$

Classification *Entropy* (Goodchild *et al.* 1994) in its normalised form is given by:

$$Entropy = \left| \frac{1}{\log_e c} \sum p_c \log_e p_c \right|. \qquad [B.2]$$

The normalised versions of the uncertainty metrics in Eq. B.1 and B.2 provide consistent uncertainty values between 0 and 1 regardless of the number of possible classes.

The differences between the normalised versions of *Variance* and *Entropy* are small but related to the slope of curves for uncertainty values proximal to 0 and 1 (Figure B.1). This figure compares *Variance* and *Entropy* decay curves for different numbers of possible classes. These curves are generated by varying the maximum p_c for one class. All other p_c are distributed equally across the remaining proportion, i.e. 1 maximum probability. Where $c > 8$, *Variance* curves increase the weight of small deviations from p_c of 0 or 1 (van der Wel *et a l.* 1998), thus emphasising slightly higher uncertainties when compared to *Entropy*.

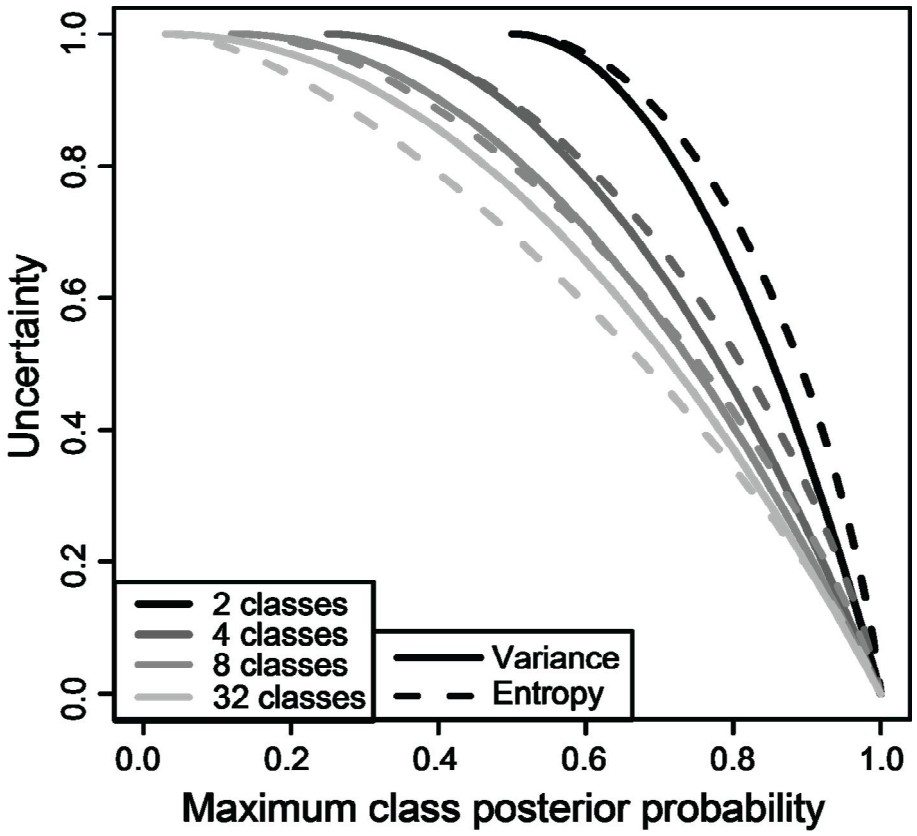

Figure B.1 Examples of Variance and Entropy decay curves for different numbers of classes. Note that for c > 8, Variance emphasises higher uncertainties than Entropy.

APPENDIX C – SUPPLEMENTARY INFORMATION

The information contained within this supplementary document concerns key aspects of Chapter 4 (Cracknell & Reading 2014) entitled "Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information" that were not included due to restrictions on article length. Summaries of the geophysical, i.e. airborne geophysics and Landsat ETM+, data used are given in Section C.1 along with detailed descriptions of the pre-processing methods employed to prepare inputs for machine learning algorithm training. In Section C.2, we provide citations for and briefly describe machine learning algorithm software used in this study. This includes information on the specific parameters that require optimisation when training machine learning algorithms for supervised classification applications.

C.1. Data

Airborne geophysics and Landsat ETM+ multispectral satellite imagery were used as inputs for machine learning algorithm (MLA) training and lithology prediction (Table C.1). Airborne geophysics was supplied as interpolated and gridded data (Australian Geological Survey Organisation, 1994), while Landsat ETM+ data was supplied as gridded data with Level 1 processing applied (National Aeronautics and Space Administration 2010). Level 1 processing includes a Payload Correction Data processing, Mirror Scan Correction Data processing, geometric correction using Ground Control Points and terrain correction (Williams 2009). Table C.2 summarises the Landsat 7 ETM+ band spectral wavelengths.

A series of pre-processing steps and transformations were applied to the original data such as: potential field processing (Total Magnetic Intensity: TMI); noise reduction (Gamma-Ray Spectrometry: GRS); and calculating band ratios for selected bands (Landsat ETM+ and GRS). Reduction-to-Pole (RTP) is a potential field processing step that accounts for non-vertical inclinations in the TMI field around magnetic features at mid-latitudes and adjusts them such that they are inclined vertically over the magnetic features they represent. The first vertical derivative (1VD) of RTP TMI was also calculated as this has the benefit of enhancing shallow magnetic features (Telford *et al.* 1990). Pre-processing of

Table C.1 Broken Hill sample area input variable summary, modified from Cracknell and Reading (2013).

Variable	Units	Resolution/scale	Pre-processing
X	Eastings m (integer)	50 m	-
Y	Northings m (integer)	50 m	-
DEM	m ASL (float)	~ 21 m	-
RTP	nT (float)	~ 21 m	-
1VD	nT (float)	~ 21 m	-
TC	count (float)	~ 21 m	negative value correction and mean focal filter (5 × 5 kernel)
K	% (float)	~ 21 m	negative value correction and mean focal filter (5 × 5 kernel)
Th	ppm (float)	~ 21 m	negative value correction and mean focal filter (5 × 5 kernel)
U	ppm (float)	~ 21 m	negative value correction and mean focal filter (5 × 5 kernel)
logK-Th	ratio (float)	~ 21 m	natural logarithm of the ratio between K and Th and mean focal filter (5 × 5 kernel)
logU-Th	ratio (float)	~ 21 m	natural logarithm of the ratio between U and Th and mean focal filter (5 × 5 kernel)
Landsat-1	DN (8 bit integer)	28.5 m	-
Landsat-2	DN (8 bit integer)	28.5 m	-
Landsat-3	DN (8 bit integer)	28.5 m	-
Landsat-4	DN (8 bit integer)	28.5 m	-
Landsat-5	DN (8 bit integer)	28.5 m	-
Landsat-6 (sensor 2)	DN (8 bit integer)	57 m	-
Landsat-7	DN (8 bit integer)	28.5 m	-
Landsat-3-1	ratio (float)	28.5 m	ratio of Landsat band 3/band 1
Landsat-3-2	ratio (float)	28.5 m	ratio of Landsat band 3/band 2
Landsat-3-5	ratio (float)	28.5 m	ratio of Landsat band 3/band 5
Landsat-3-7	ratio (float)	28.5 m	ratio of Landsat band 3/band 7
Landsat-5-1	ratio (float)	28.5 m	ratio of Landsat band 5/band 1
Landsat-5-2	ratio (float)	28.5 m	ratio of Landsat band 5/band 2
Landsat-5-4	ratio (float)	28.5 m	ratio of Landsat band 5/band 4
Landsat-5-7	ratio (float)	28.5 m	ratio of Landsat band 5/band 7
Landsat-5-4x3-4	ratio (float)	28.5 m	ratio of Landsat band 5/band 4 × band 3/band 4
Geology	categorical (text)	1: 250,000	rasterised to 50 m pixels

Table C.2 Landsat 7 ETM+ spectral bands and bandwidths, based on Williams (2009).

Spectral Bands		Half-Amplitude Bandwidth (μm)
1	Blue-Green	0.450–0.515 ± 0.005
2	Green	0.525–0.605 ± 0.005
3	Red	0.630–0.690 ± 0.005
4	Near Infra-Red	0.775–0.900 ± 0.005
5	Short Wave Infra-Red	1.550–1.750 ± 0.010
6	Thermal Infra-Red	10.40–12.50 ± 0.100
7	Short Wave Infra-Red	2.090–2.350 ± 0.020
8 (Panchromatic)	Visible to Near Infra-Red	0.520–0.900 ± 0.010

TMI data was conducted using the geophysical processing tools available in ERDAS ERMapper Version 7.2. The declination and inclination parameters for RTP processing were obtained from the Australian Geomagnetic Reference Field (Geoscience Australia 2010). GRS negative values (instrument drop out) in all four channels were replaced with a positive value close to zero but an order of magnitude less than the lowest positive value, i.e. 0.001, in for each survey. Corrected GRS bands were smoothed using a 5 × 5 mean focal filter. Dauth (1997) found that discriminating between fresh and weathered mafic bedrock is assisted using the K/Th band ratio. Whereas, Gabriel (2007) identified a correlation between fractionation zoning in felsic igneous rocks could be identified by using the U/Th band ratio. Based on this information, Potassium (K), Uranium (U) and Thorium (Th) bands were further processed by the calculation of ratios K/Th and U/Th. These data were normalised prior to generating band rations by taking the natural logarithm.

Landsat band ratios are useful for reducing the effect of shadows generated from sun angle illumination on topography (Congalton & Green 1998; Durning *et al.* 1998). Several Landsat band ratios deemed to be beneficial for the discrimination of geological and/or soil materials were calculated and used for training and lithology prediction. Landsat band ratios 5/2 and 3/5 both assist in distinguishing between calcareous sedimentary and mafic igneous rocks (Boettinger *et al.* 2008; Mshiu 2011). Band ratios, 5/1 and 5/4 × 3/4 were found to be useful for distinguishing volcanic and metamorphic rocks from sedimentary rocks (Kusky & Ramadan 2002). Landsat band ratios are useful for discriminating areas containing 3/1 ferric iron, 3/2 carbonate rocks and 5/7 hydroxyl ions associated with clays and alteration (Amen & Blaszczynski, 2001 in Durning *et al.* 1998; Inzana *et al.* 2003; Boettinger *et al.* 2008, p. 197). Both band ratios 3/7 (Amen & Blaszczynski, 2001 in Boettinger *et al.* 2008, p. 197) and 5/4 (Durning *et al.* 1998) have been found useful for identifying ferrous iron.

All acquired data inputs covering the Broken Hill area were aggregated prior to resampling to a common spatial domain. Aggregation merges neighbouring pixels into a larger pixel using a function of the value of these pixels, such as the mean, to set the resulting pixel value (Hijmans & van Etten 2012). Landsat bands were aggregated from 28.5 m to 57 m and resampled to 50 m, except for Landsat band 6 (sensor 2) which is already available at 57 m resolution and thus only required resampling. Broken Hill airborne geophysics data, originally obtained with ~ 21 m pixel dimensions were aggregated to ~ 42 m and then

resampled to 50 m. Samples representing target classes *y* were obtained by rasterising the digital geological maps with the same extent and pixel sizes as defined for the resampled input layers.

C.2. MLA software and parameters

The R programming language, version 2.15.0 (64-bit), available from The Comprehensive R Archive Network (http://cran.r-project.org/), was used exclusively to implement and evaluate MLAs. Functions for input data selection, MLA training and evaluation of classification model outputs relied heavily on the *caret* (Kuhn *et al.* 2012) and *raster* (Hijmans & van Etten 2012) packages. Parametric input importance rankings were constructed using the *caret* package which employs *pROC* (Robin *et al.* 2011) to evaluate relative individual variable importance. MLA parameters were selected using 10-fold cross-validation.

For Naïve Bayes (NB), single classification model parameter, *usekernel*, requires selection. This parameter adjusts the way in which class conditional densities are estimated; *usekernel* = FALSE assumes that the marginal distributions are normal. Alternatively, *usekernel* = TRUE implements a non-parametric kernel density estimation technique to establish priors for each input (John & Langley 1995). NB was implemented using *klaR* (Weihs *et al.* 2005), which is based on the nb() function in the *e1071* package (Dimitriadou *et al.* 2011).

The *k*-Nearest Neighbours (kNN) algorithm requires *k* to be selected, representing the number of nearest neighbours to compare in the training dataset. kNN training and prediction was carried out using the predknn() function in the *ipred* package (Peters & Hothorn 2011).

Random Forests (RF) requires the selection of two tuning parameters, *trees* and *mtry*. The number of random decision trees to construct for each forest is set by *trees* while the number of randomly selected layers to split at the nodes of each tree is set by *mtry*. We have maintained *trees* at the default value of 500 and used cross-validation to select an optimal *mtry* value. The randomForest() function from the *randomForest* package (Liaw & Wiener 2002) was used to train this algorithm.

Although there are many types of kernels that can be used with Support Vector Machines (SVM), the Radial Basis Function was used in this experiment as is it a good first choice and the default for the kernlab() function in the *kernlab* package (Karatzoglou *et al.* 2004). In conjunction with the kernel function, two other parameters require selection when implementing SVM, ó and *C*. The ó parameter sets the width of the kernel function and *C* adjusts the sensitivity of the decision margin to misclassified support vectors. We estimated ó using the sigest() function *kernlab* while cross-validation was used select optimal *C* values.

We have explored one of the two possible tuning parameters available for the Multi-Layer Perceptron (MLP) Artificial Neural Network (ANN) used in this study. This parameter, *size*, sets the number of hidden layers in the neural network. Increasing the number of hidden layers increases the complexity of the resulting network. We have maintained the other possible ANN parameter *decay*, at the default value of 0. Previous experimentation indicated that *size* influenced ANN prediction accuracies far more than variations in the *decay* parameter. ANN was implemented using the nnet() function available in *nnet* package (Venables & Ripley 2002).

APPENDIX D – R PACKAGES

R version 2.15.0 64-bit was used to conduct the analysis performed as part of the study documented in Chapter 5 (Cracknell & Reading 2013). General source packages and functions for R are available from the Comprehensive R Archive Network (http://cran.r-project.org/). Spatial vector (points, lines and polygons) data management, processing and analysis was conducted using *sp* (Pebesma & Bivand 2005; Bivand *et al.* 2008), *maptools* (Lewin-Koh *et al.* 2012), *rgdal* (Keitt *et al.* 2012) and *rgeos* (Bivand & Rundel 2012). Spatial raster (image/regular array) data management, processing and analysis were carried out using *raster* (Hijmans & van Etten 2012). The Euclidian distance from lithology boundaries were calculated using *spatstat* (Baddeley & Turner 2005). Methods for data pre-processing, machine learning classification model training and parameter selection were sourced from *caret* (Kuhn *et al.* 2012). Random Forests was implemented using *randomForest* (Liaw & Wiener 2002) and Support Vector Machines using *kernlab* (Karatzoglou *et al.* 2004).

APPENDIX E – DATA SOURCES AND PRE-PROCESSING

Prior to training Random Forests for the study documented in Chapter 6 (Cracknell *et al.* 2014), classes in the unpublished geological map interpreted by Richardson (1994) with < 300 samples were incorporated into surrounding units or units with similar lithological characteristics. Intrusive bodies of andesite volcaniclastic rocks (ADi) were merged with the feldspar-phyric andesite (Afp). Barite (Ba) was merged with strongly altered rock (HA). Dacite intrusions (Dvc) were incorporated into the surrounding Animal Creek Greywacke (ACG) unit. A small Quaternary glacial deposit (Qg) was merged into the underlying body of Afp.

Airborne geophysical data used in this study originated from several different sources. High resolution Total Magnetic Intensity (TMI) and Gamma-Ray Spectrometry (GRS) data were obtained from the 1993 Mackintosh geophysical survey conducted by Aberfoyle Resources Limited. The flight line spacing (100 m) and ground clearance (~ 80 m) specifications of this survey (Richardson 1993) were designed to target surface/near-surface geological features. These data are publicly available from Mineral Resources Tasmania (http://www.mrt.tas.gov.au) as an interpolated and levelled raster file at 20 m resolution.

TMI data were Reduced-To-Pole (RTP) in ERDAS ERMapper 7.2 using parameters (dec. = 12.376, inc. = 71.887) calculated from the Australian Geomagnetic Reference Field (Geoscience Australia 2010). The 1st Vertical Derivative of the RTP residual field (1VD) was calculated using ERDAS ERMapper 7.2. The RTP TMI data contains a small but significant line of artefacts in the west of the study region indicating high-voltage power lines. The effects of these spurious features were reduced by smoothing using a mean 5 × 5 convolution filter. An observable regional magnetic gradient in the RTP TMI data (Richardson 1994) was removed by subtracting a linear trend surface. GRS data were checked to ensure that negative values were not present and smoothed to reduce the effects of noisy observations via a mean 5 × 5 convolution filter. GRS band ratios K/Th, U/Th and U^2/Th were calculated and included as input variables.

The Digital Elevation Model (DEM) and Airborne Electro-Magnetics (AEM) data form part of the Western Tasmanian Regional Minerals Project survey conducted by the Tasmanian Geological Survey in 2002. East-west flight line spacing for the AEM survey was 200 m and the nominal sensor height was ~ 30 m (Reid 2003). AEM data are available at 40 m pixel resolution. DEM (m) and $\log_{10}$ AEM apparent resistivity (ohm-m) data were smoothed using a mean 5 × 5 convolution filter to reduce the presence of outliers and dampen artefacts due to high-voltage power lines. Five different components of the $\log_{10}$ AEM apparent resistivity maps were provided in the source data. These relate to the different frequencies of data acquisition: 34,111 Hz; 7004 Hz; 6600 Hz; 985 Hz; and 880 Hz. Apparent resistivity maps were calculated using the amplitude-altitude algorithm, which assumes a perfectly conductive (non-magnetic) Earth (Huang & Fraser 2000). An accurate knowledge of sensor altitude is required for the amplitude-altitude algorithm, which for radar altitude measurements can be difficult to accurately estimate in densely vegetated areas due to returns from the canopy (Reid 2003).

Soil geochemical data were collated by Aberfoyle Resources Limited and contain ~ 26,000 separate georeferenced points attributed with the analysed abundance of 11 trace elements (Cu, Pb, Zn, Ag, Au, Ba, As, Cr, Zr, Ti and Ni). These samples, assumed to relate to C-horizon soil profiles, were collected by hand auger to a depth of 1 m. Mean elemental abundances were calculated for samples within 5 m of each other. The Au dataset contained less than 200 samples and was not included in this study.

Rigorous Quality Assurance/Quality Control (QA/QC) procedures ensured comparable geochemical assay data between different samples. QA/QC was based on laboratory internal procedures, internal standards and repeat sampling. All soil geochemical samples were air-dried (~ 30 °C) then sieved with the 80# fraction analysed. A single standard was used for all sample batches. This standard was derived from one crushed, pulverised and homogenised andesite drill hole intersection. This standard was submitted as at least 1 standard per 50 samples.

The interpolation of geochemical data from irregularly (spatially) distributed points to a uniform grid with 20 m pixel resolution was carried out using kriging with fixed covariance parameters for a global neighbourhood. Covariance parameters were estimated by weighted least-squares fitting of a parametric model to an isotropic empirical variogram for distances of 3 km or 5 km at 100 m intervals. Exponential or spherical variogram

models were fitted to these data. A maximum of 10,000 randomly sampled points was selected for variogram fitting and kriging. As the majority of samples contained missing assay values, random sampling only reduced the number of samples available to interpolate Cu, Pb and Zn data. Despite Cu, Pb and Zn being important geochemical elements for targeting volcanic-hosted massive sulfide deposits, the interpolation results for these elements were not significantly affected by random sampling because of the high spatial densities of samples with respect to the output pixel resolution. The $\log_e$ Ti/Zr ratio was calculated and included in the classification procedure as it is an important measure for discriminating among QHV units (Corbett & Komyshan 1989; Crawford *et al.* 1992; Gemmell & Fulton 2001).

Landsat ETM+ spectral radiance data (NASA 2002), publically available from the United States Geological Survey (http://eros.usgs.gov), were not affected by cloud cover over the study area. Furthermore, the region under investigation is small enough (~ 36 km^2) such that large scale atmospheric effects would not have varying impacts on different regions of the scene. Therefore, we assume that the radiance of imaged surface materials and objects of interest display sufficiently different reflectance characteristics in order to differentiate them and that atmospheric effects are minor enough to not affect their spectral separation.

Landsat ETM+ data were supplied as multiple bands with Level 1 processing, which includes a Payload Correction processing, Mirror Scan Correction processing, geometric correction using Ground Control Points and terrain correction (Williams 2009). From the seven original bands in the Landsat ETM+ data, ratios of bands 3/1, 3/2, 3/5, 3/7, 5/1, 5/2, 5/4, 5/7 and 5/4 x 3/4 were calculated. These band ratios have been used in previous research to discriminate geological features, including hydrothermally altered zones and weathering products (e.g., Durning *et al.* 1998; Kusky & Ramadan 2002; Inzana *et al.* 2003; Boettinger *et al.* 2008; Mshiu 2011).

APPENDIX F – R CODE AND SCRIPTS

Appendix F is in digital format and supplied as data and files on the USB flash drive that accompanies this thesis. A copy of the README.txt file associated with Appendix F is provided below. A PDF version of this thesis can also be found on the USB flash drive.

README.txt

12 December 2013

M. J. Cracknell, m.j.cracknell@utas.edu.au

The accompanying files contain R code and scripts (and associated data) developed and documented in the PhD thesis entitled "Machine Learning for Geological Mapping: Algorithms and Applications". These scripts implement machine learning algorithms (MLAs) for remote sensing 2D geological mapping applications. The R programming language is freely available from the Comprehensive R Archive Network (http://cran.r-project.org/). All code was written and executed using R version 2.15.0 64-bit (2012-03-30) copyright (c) 2012 the R Foundation for Statistical Computing using a DELL Desktop PC 64-bit Intel Dual Core CPU @ 3.00 GHz and 8 GB of RAM.

The directory structure for Appendix F is as follows:

- Digital_Appendix_F\\
    - Data\\
        - BrokenHill_comparison\\ – Chapter 4 data and files
        - BrokenHill_uncertainty\\ – Chapter 5 data and files
        - Hellyer\\ – Chapter 6 data and files
        - Rosebery_spatial-context\\ – Chapter 7 data and files
    - Rcode\\ – R code and scripts as summarised in Table 8.1 (p. 200)

Save the Digital_Appendix_F folder onto the C:\\ drive as the path names in the scripts link to this location. Although every effort has been made to ensure that these R scripts and data are free from bugs and errors, however, users may have issues with more recent versions of R (> 2.15.0). For more information and/or assistance using these R scripts please contact the author.